3D ENANTIOSELECTIVE DESCRIPTORS FOR LIGAND-BASED COMPUTER-AIDED DRUG DESIGN

By

Gregory Richard Sliwoski

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Interdisciplinary Studies: Interdisciplinary Pharmacology

August, 2012

Nashville, Tennessee

Approved:

Professor Jens Meiler, Ph.D.

Professor Charles David Weaver, Ph.D.

Professor Vsevolod V Gurevich, Ph.D.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER I

LIGAND-BASED COMPUTER AIDED DRUG DESIGN

Introduction

On October 5, 1981, Fortune magazine published a cover article entitled the "Next Industrial Revolution: Designing Drugs by Computer at Merck" [1]. Some have credited this as being the start of intense interest in the potential for Computer Aided Drug Design (CADD). While progress was being made in CADD, the potential for high-throughput screening (HTS) had begun to take precedence as a means for finding novel therapeutics. This brute force approach relies on automation to screen high numbers of molecules in search of those which elicit the desired biological response. This method requires little compound design or prior knowledge and the efficiency of technologies required to screen large libraries continues to increase. However, while traditional HTS is often successful in the discovery of multiple lead compounds, the hit rate for this method is extremely low. This low hit rate has limited the usage of HTS to research programs capable of screening very large compound libraries. In the past decade, CADD has reemerged as a way to significantly decrease the number of compounds necessary to screen while retaining the same level of lead compound discovery. CADD techniques allow compounds predicted as inactive to be skipped and those predicted as active to be prioritized. This reduces the cost and workload of a full HTS screen without sacrificing lead discovery. For example, researchers at Pharmacia (now part of Pfizer) used CADD tools to screen for inhibitors of tyrosine phosphatase-1B, an enzyme implicated in diabetes. Their CADD-based virtual screen yielded 365 compounds, 127 of which showed effective inhibition, a hit rate of nearly 35%. Simultaneously, this group performed a traditional HTS against the same target. Of the 400,000

compounds tested, 81 showed inhibition, producing a hit rate of only .021%. This comparative case effectively displays the power of CADD for reducing the number of compounds necessary to test for hit discover [2]. CADD has already been used in the discovery of compounds which have passed clinical trials and become novel therapeutics in use for the treatment of a variety of diseases. Some of the earliest examples of approved drugs that owe their discovery in large part to the tools of CADD include the carbonic anhydrase inhibitor dorzolamide, approved in 1995 [3], the ACE inhibitor captopril, approved in 1981 as an antihypertensive drug [4], three therapeutics for the treatment of HIV: saquinavir (approved in 1995), ritonavir and indinavir (both approved in 1996) [1] and tirofiban, a fibrinogen antagonist approved in 1998 [5].

One example that helps validate the use of CADD in lead compound discovery is the search for novel TGF-beta-1 receptor kinase inhibitors in 2003. One group at Eli Lilly used a traditional high throughput screening to identify a lead compound that was subsequently optimized [6], while a group at Biogen Idec used a CADD approach involving virtual HTS based on the structural interactions between a weak inhibitor and TGF-beta-1 receptor kinase [7]. Upon the virtual screening of compounds, the group at Biogen Idec identified 87 hits, the best hit being identical in structure to the lead compound discovered through the traditional HTS approach at Eli Lilly [8]. In this situation CADD, a method involving reduced cost and workload, was capable of producing the same lead as a full-scale HTS.

Position of CADD in the drug discovery pipeline

CADD is capable of increasing the hit rate of novel drug compounds as it employs a much more targeted search than traditional HTS and combinatorial chemistry. It not only aims to explain the molecular basis of therapeutic activity, but also to predict possible derivatives that would improve activity. One of the most common uses in CADD is the screening of virtual

compound libraries, also known as virtual high-throughput screening (vHTS). This allows experimentalists to focus resources on testing compounds likely to have an activity of interest. Ripphausen et al. note that the first mention of vHTS was in 1997 [9] and chart an increasing rate of publication for the application of vHTS between 1997 and 2010. They also found that the largest fraction of hits has been obtained for GPCR's, followed by kinases [10].

vHTS comes in many forms including chemical similarity searches by fingerprints or topology, selecting compounds by predicted biological activity through Quantitative Structure-Activity Relationship (QSAR) models or pharmacophore mapping, and structure-based docking [11]. These methods allow the ranking of "hits" from the virtual compound library for acquisition. The ranking can reflect a property of interest such as percent similarity to a query compound, predicted biological activity, or in the case of docking, the lowest energy scoring poses for each ligand bound to the target of interest. Often initial hits are rescored and ranked using higher level computational techniques that are too time-consuming to be applied to full scale vHTS. It is important to note that vHTS does not aim to identify a drug-compound that is ready for clinical testing, but rather to find leads with chemotypes that have not previously been associated with a target. This is not unlike a traditional HTS. Through iterative rounds of chemical synthesis and *in vitro* testing, a compound is developed into a "lead" with higher affinity and some understanding of is structure-activity-relation. This lead can then be tested for its DMPK/ADMET properties. Only after further iterative rounds of lead-to-drug optimization and *in vivo* testing does a compound reach a clinically appropriate potency and acceptable DMPK/ADMET properties [12]. For example, the literature survey performed by Ripphausen et al revealed that a majority of successful vHTS applications identified a small number of hits that are active in the micromolar range, and hits with low nanomolar potency are only rarely identified [10].

Ligand-Based Computer-Aided Drug Design (LB-CADD)

The ligand-based computer-aided drug discovery (LB-CADD) approach involves the analysis of ligands known to interact with a target of interest. These methods utilize a set of reference structures collected from compounds known to interact with the target of interest and analyze their 2D or 3D structures. The overall goal is to represent these compounds in such a way that the physicochemical properties most important for their desired interactions are retained while extraneous information not relevant to their interactions is discarded. LB-CADD is based on the Similar Property Principle, published by Johnson and Maggiora, which states that molecules that are structurally similar are likely to have similar properties [13]. It is considered an indirect approach to drug discovery in that it does not necessitate any prior knowledge of the target of interest. LB-CADD approaches are commonly applied when the 3D structure of the biological target is unknown. The two fundamental approaches of LB-CADD are a) selection of compounds based on chemical similarity to known actives using some similarity measure or b) the construction of a Quantitative Structure-Activity Relation (QSAR) model that predicts biological activity from chemical structure. The difference between the two approaches is that the latter weights features of the chemical structure according to their influence on the biological activity of interest, the former does not. The methods are applied for *in silico* screening for novel compounds possessing the biological activity of interest, hit-to-lead and lead-to drug optimization, and also for the optimization of DMPK/ADMET properties.

Molecular Fingerprint and Similarity Searching

Molecular fingerprint-based techniques attempt to represent molecules in such a way as to allow rapid structural comparison in an effort to identify structurally similar molecules or

to cluster collections based on structural similarity. These methods are less hypothesis-driven and less computationally expensive than pharmacophore mapping or QSAR. They rely entirely on chemical structure and do not take compound potency/activity into account, making the approach more qualitative in nature than other LB-CADD approaches [14]. Additionally, fingerprint-based methods do not attempt to focus only on parts of a molecule that are thought to be most important for activity but rather they consider all parts of the molecule equally. This is less prone to errors than hypothesis-driven methods but suffers from the influence of unnecessary features [14]. Despite this drawback, 2-dimensional fingerprints continue to be the representation of choice for similarity-based virtual screening [15]. Not only are these methods the computationally least expensive way to compare molecular structures [16], but their effectiveness has been demonstrated in many comparative studies [15].

Fingerprint Types

Fingerprints are bit string representations of molecular structure and/or properties [17, 18]. They encode various molecular features as pre-defined bit settings [14] i.e. representation as 1 or 0, where 1 means feature is present or 0 if not. This allows chemical identity to be unambiguously assigned by the presence or absence of specific features [16]. The features described in a molecular fingerprint can vary in number and complexity (from hundreds of bits for structural fragments to thousands for connectivity fingerprints, and millions for the complex pharmacophore-like fingerprints) [14], depending on the computational resources available and the intended application. Fingerprints which rely solely on interatomic connectivity – i.e. molecular constitution – are known as 2-dimensional fingerprints [16]. In the prototypic 2D keyed fingerprint design, each bit position is associated with the presence or absence of a specific substructure pattern – for example carbonyl group attached to sp^3 carbon, hydroxyl group attached to sp^3 carbon, etc. [19].

Molecular structure itself comprises several levels of organization between the atoms within a molecule and therefore fingerprints too may differ in their levels of organization. For example, the simplest fingerprint may contain the information that a given compound contains six carbon atoms and six hydrogen atoms. However, up to 217 different isomers can contain this fingerprint. Adding connectivity increases the specificity of the fingerprints but does not necessarily provide discrimination between stereoisomers. These molecules are not identical despite have equal constitutions and 2D fingerprints are insufficient to distinguish their structures. Therefore, considerable effort is taken to ensure the efficient application of fingerprints without sacrificing important molecular characteristics. One extension to fingerprints is the use of hash codes. These are bit strings of fixed length that contain information about connectivity, stereocenters, isotope labeling, and further properties. This information is then compressed to avoid redundancies [20]. Unfortunately, it is not always obvious which of these characteristics are important in a given context and which are not [16].

Commonly used bit strings include the ISIS (Integrated Scientific Information System) keys with 166 bits and the MDL (Molecular Design Limited) [21] MACCS (Molecular ACCess System) keys [22] with 960 bits. The ISIS keys are small topological substructure fragments while the MACCS keys consist of the ISIS keys plus algorithmically generated more abstract atom-pair descriptors. MDL keys are commonly used when optimizing diversity [23, 24]. For example, the PubChem database uses a fingerprint that is 881 bits long to rank substances against a query compound. This fingerprint is comprised of the number and type of elements, ring systems (saturated and unsaturated up to a size of 10), pair-wise atom combinations, sequences, and substructures [16].

Similarity Measures

Molecular fingerprints are commonly used in ligand-based drug design to search a large database of fingerprints against a query molecule. In order to be effective, the search algorithm must employ a means of scoring the similarity between the query fingerprint and those in the database. Pairwise comparison of bit string overlap serves as the criterion for similarity and is based on the calculation of similarity coefficients [25]. The most commonly used similarity coefficient is the Tanimoto coefficient and is defined by the equation:

$$Tanimotto\ Coefficient\ = \frac{NAB}{NA\ + NB - NAB}$$

where NA is the number of bits set to 1 in fingerprint A, NB is the number of bits set to 1 in fingerprint B, and NAB is the number of common bits [14]. The Tanimoto coefficient, however, is not always the best similarity coefficient. For example, it typically yields low similarity values when the query fingerprint has just a few bits set to 1 [26].

While the Tanimoto coefficient is the most popular score equation, there are several others that have been proposed and the use of different scoring coefficients has been explored. Willett, et al. compares the different coefficients in detail, considering the use of data fusion methods for combining the results of database searches that use the same query but different similarity coefficients. They were unable to identify a single combination of coefficients that yielded the greatest performance in all circumstances. However, they did find that different coefficients consistently performed better than others within a given molecular size. For example, the Russell-Rao coefficient appeared in many of the best combinations involving smaller active molecules and the Tanimoto coefficient tended to retrieve molecules from the center of the size distribution[15].

Similarity Searches in LB-CADD

Fingerprint methods may be employed to search databases for compounds similar in structure to a lead query, providing an extended collection of compounds that can be tested for improved properties over the lead. In many situations, 2D similarity searches of databases are performed using chemotype information from first generation hits, leading to modifications that can be evaluated computationally or ordered for in vitro testing [4]. Bologa et al used 2D fingerprint and 3D shape-similarity searches to identify novel agonists of the estradiol receptor family receptor GPR30. This work yielded a first-in-class selective agonist with a K_I of 11 nM [27]. SecinH3, a lead compound targeting cytohesins involved in insulin signaling initially identified with classical HTS, was used as a query molecule in a 2D-fingerprint search that yielded 26 novel cytohesin inhibitors, all of which were more potent than SecinH3 [28]. 2D pharmacophoric fingerprints were also used to identify novel T-type calcium channel blockers. Of the 38 molecules selected for testing, 16 showed more than 50% blockade of CaV3.2 mediated T-type current. These compounds proved to be an interesting collection of T-type calcium channel blockers. Some showed reversible inhibition while others resulted in irreversible inhibition and one of the compounds caused alterations in depolarization/repolarization kinetics [29].

In addition to the enrichment of lead compound population, fingerprints are also used to increase molecular diversity of test compounds. Fingerprints can be used to cluster large libraries of hits in order to allow the sampling of a wide range of compounds without the need to sample the entire library. In this case, fingerprints are used to optimize the sampling of diversity space. The Jarvis-Patrick method which calculates a list of nearest neighbors for each molecule has been shown to perform well for chemical clustering. Two structures cluster together if they are in each-others list of nearest neighbors and they have at least K of their J

nearest neighbors in common. The MDL keys also provide a way to eliminate compounds which are least likely to satisfy the drug-likeness criterion [23].

Fingerprint Extensions

Current research is focused on improving fingerprint-based LB-CADD methods. As mentioned, one drawback with fingerprint-based methods is that all features of a query molecule are equally important for ranking candidate molecules, regardless of any effect of these features on the biological activity at a target. One group, Hessler et al. proposes a method that intends to combine the advantages of similarity and pharmacophore searching on the basis of 2D structural representations only. In their proposed method, a set of query molecules is converted into a topological model (MTree) based on chemically reasonable matching of corresponding functional groups. This creates a topological map of the most similar fragments from a set of structurally diverse but active molecules and conserved features are characterized by high similarity scores of the corresponding nodes in the MTree model [30]. Due to the low dependence on chemical substructures, they argue that the MTree model is especially useful for identification of alternative novel molecular scaffolds or chemotypes. Methods for forming multiple feature tree models and multiple feature tree scoring schemes are also presented.

Pharmacophores – Superimposing Active Compounds

In 1998, the IUPAC formally defined a pharmacophore as 'the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response' [31]. In terms of drug activity, a pharmacophore is the spatial arrangement of functional groups that a compound or drug must contain in order to evoke a desired biological response. Therefore, an effective pharmacophore will contain information about functional groups that interact with the

target, as well as information regarding the type of non-covalent interactions and interatomic distances between these functional groups/interactions. This arrangement can be derived either in a structure-based manner by mapping the sites of contact between a ligand and binding site, or using a ligand-based approach. To generate a ligand-based pharmacophore, multiple active compounds are overlaid in such a way that a maximum number of chemical features overlap geometrically [32]. This can involve rigid 2D or 3D structural representations or, in more precise applications, incorporate molecular flexibility to determine overlapping sites. This conformational flexibility can be incorporated by pre-computing the conformational space of each ligand and creating a general-purpose conformational model or conformations can be explored by changing molecule coordinates as needed by the alignment algorithm [32]. For example, one popular pharmacophore-generating software package, Catalyst, uses the Poling algorithm [33] to generate approximately 250 conformers that it uses in its pharmacophore generation algorithm [34]. In a study targeting HSP90α, Al-Sha'er *et al* used 83 known reference molecules to generate pharmacophore queries and identified twenty-five diverse inhibitors including three with IC_{50} values below 10 nM [35].

Pharmacophore Overlaying

Molecules are commonly aligned through either a point-based or property-based technique. The point-based technique is the most widely used method and involves superposing pairs of points (atoms or chemical features) by minimizing Euclidean distances. The alignment of pharmacophore features is the most commonly used method when screening libraries against a query pharmacophore as well as for generating a pharmacophore. Property-based alignment techniques, on the other hand, use molecular field descriptors to generate alignments. They define a grid around each ligand and calculate interaction energies at each point between the

ligand and specific probe molecules. Overlapping interaction energies are used to guide alignments [32].

Pharmacophore feature extraction

A pharmacophore feature map is carefully constructed so as to balance generalizability with specificity. A general definition might categorize all functional groups having similar physiochemical properties (i.e. similar hydrogen-bonding behavior, ionizability) into one group whereas specific feature definitions may include such things as specific atom types at specific locations. More general feature definitions allow the identification of novel scaffolds and increase the population of compounds that match the pharmacophore. However, some degree of restriction is necessary for a pharmacophore's predictive power to avoid high numbers of false positives which would result in poor LB-vHTS performance. The level of feature definition generalizability is usually determined by the algorithm used to extract feature maps and through user-specified parameters. The most common features used to define pharmacophore maps are hydrogen bond acceptors and donors (covalently bound, partially positive hydrogen atoms interact with a partially negative atom), acidic and basic groups (groups of atoms that are likely to be protonated or deprotonated at physiological pH), aromatic rings, aliphatic hydrophobic moieties, and aromatic hydrophobic moieties [34]. These are commonly implemented as spheres with a certain tolerance radius for pharmacophore matching [32].

Pharmacophore Algorithms and Software Packages

The most common software packages employed for ligand-based pharmacophore generation include Phase [36], MOE [37], Catalyst [38, 39], and LigandScout [40]. Catalyst contains multiple tools for constructing pharmacophores. One such tool, HipHop, generates pharmacophores based on active molecules. Catalyst HypoGen is another pharmacophore module of Catalyst that uses a full range of training set compounds including inactive and active

compounds. The pharmacophore generated reflects not only features common among active compounds but also features missing from inactive compounds. HypoGen first finds all common features amongst actives and then removes the features common amongst the inactives. Simulated annealing is then employed to optimize the results returning up to ten different models [41]. These software packages provide different strengths and weaknesses depending on different implementations of certain features. For example, Catalyst only permits a single bonding feature per heavy atom while LigandScout allows a hydrogen-bond donor or acceptor to be involved in more than one hydrogen-bonding interaction [32]. MOE, on the other hand, allows a more customizable approach to hydrogen-bonding features. Lipophilic areas are generally represented as spheres located on hydrophobic atom chains, branches, or groups in a similar manner across software packages but with slight nuances. While subtle, these differences have important consequences on prediction models. For example, software packages that do not attach a hydrophobic feature to an aromatic ring are unable to predict that an aromatic group may be positioned in a lipophilic binding pocket [32]. Different algorithms provide different levels of customizability for user defined features. Catalyst allows the specification of one or more chemical groups that satisfy a particular feature while Phase allows not only matching chemical groups but also a list of exclusions for a given feature. MOE offers a level of customization that allows the user to implement entirely novel pharmacophore schemes as well as modification of existing schemes. However, this requires additional levels of expertise to program[32]. DISCO is another commonly used pharmacophore tool that determines the spatial orientation of common points among all active compounds and incorporates flexibility by accepting conformational ensembles for each compound. The features identified with DISCO include hydrophobic centers, hydrogen-bonding, and positive and negative charges [41]. GASP (Genetic Algorithm Similarity Program) incorporates conformational flexibility while overlaying

the compounds. It attempts to optimize the conformation by fitting them to similarity constraints and weighing the conformations that fit these constraints more than conformations that do not [41]. For a comprehensive analysis of the differences between commercial pharmacophore software packages, please see the 2008 review by Wolber et al [32].

Pharmacophore Mapping Applications

Ligand-based pharmacophore methods have been used for the discovery of novel compounds across a variety of targets, resulting in the discovery of compounds showing activity in the micromolar and nanomolar range as well as compounds that reflect proof of concept with *in vivo* disease models. Al-Sha'er *et al* used 83 known Hsp90-α inhibitors to generate a pharmacophore model which resulted in the identification of several compounds, including one with an IC_{50} of 25 nM [35]. Schuster, *et al*. used Catalyst to create a pharmacophore model that was used to screen for 17β-HSD3 inhibitors. Hydroxysteroid dehydrogenases (HSD3) catalyze the reduction of alcohols or carbonyls and are suggested therapeutic targets for control of estrogen and androgen-dependent diseases such as breast and prostate cancer, acne, and hair loss [42]. Fifteen top scoring hits were tested *in vitro* at 2 μM and the most potent compound was able to inhibit 17β-HSD3 by 67.1% at 2 μM [42]. Noha *et al* developed 5-point pharmacophore models using the HipHop algorithm of Catalyst based on a training set of compounds with $IC_{50} < 100$ nM against IKK-β as potential anti-inflammatory and chemosensitizing agents. The authors used 128 active and 44 inactive compounds to develop a pharmacophore model [43]. Their model was further refined with exclusion volume spheres and shape constraints to improve the scoring of compounds in their virtual high-throughput screen against the National Cancer Institute molecular database. Ten compounds were selected and the most potent compound (NSC719177) showed inhibitory activity against IKK-β in a cell free in vitro assay with IC_{50} of 6.95 μM. Additionally, this compound inhibited NF-kappaB activation induced by TNF-alpha in

HEK293 cells with an IC_{50} of 5.85 μM [43]. Chiang et al used the HypoGen module of Catalyst to generate pharmacophore models based on an indole series of 21 compounds that showed anti-proliferative activity through the inhibition of tubulin polymerization/microtubule depolymerization as novel treatments for cancer [44]. 130,000 compounds were screened and four novel compounds were discovered with anti-proliferative activity. The most potent compound displayed anti-proliferative activity in human cancer KB cells with an IC_{50} of 187 nM. This compound also inhibited the proliferation of other cancer cell types including MCF-7, NCI-H460, and SF-268 and demonstrated anti-cancer effects in a histoculture system. In vitro assays revealed that this compound inhibited tubulin polymerization with an IC_{50} of 4.4 μM [44]. Doddareddy et al generated a pharmacophore model containing 3 hydrophobic regions, one positive ionizable center, and 2 hydrogen bond acceptor groups for the identification of novel selective T-type calcium channel blockers. The most potent hit showed an IC_{50} of 100 nM [45, 46]. T-type calcium channels are involved in rhythmical firing patterns in the CNS and present therapeutic targets for the treatment of epilepsy and neuropathetic pain [29]. Manetti et al screened the Asinex and Chembridge databases using a pharmacophore designed to bind the ATP binding site of Abl. The most potent compound tested in vitro showed an IC_{50} of 16 nM [47]. Lanier et al used 3D pharmacophores containing five feature points and an exclusion sphere generated in MOE to filter a set of generated structures for optimal side chain selection for gonadotropin releasing hormone receptor [48]. 13 molecules were tested and the most active molecule showed a K_I of 50 nM. Antagonists of the H3 histamine receptor have been suggested as potential therapeutics for the treatment of obesity. Roche et al used known H3 antagonists to generate a 3D pharmacophore model with four features including a distal positive charge, an electron rich position, a central aromatic ring, and either a second basic amine or another aromatic[49]. This model was used in a *de novo* approach with the Skelgen software [50] to

generate novel compounds from fragment libraries that match the pharmacophoric restraints. Discovered compounds showed selectivity for H3 versus the other histamine receptors H1, H2, and H4. Their most potent compound showed inverse agonist activity with an EC_{50} of 200 pM in a GTPγS functional assay and a binding affinity K_I towards H3 of 9.8 nM[49].

Chao et al used pharmacophore-based design to take advantage of the therapeutic benefits of Indole-3-carbinol (I3C) in the treatment of cancer. I3C is known to suppress proliferation and induce apoptosis of various cancer cells through the inhibition of Akt activation [51, 52]. I3C, however, has a poor metabolic profile and low potency, likely due to the fact that its therapeutic behavior comes from only four of its metabolites. By overlaying these low energy conformers of these four metabolites, Chao et al was able to identify similar N-N' distances and overlapping indole rings [53]. This led them to design SR13650 which showed an IC_{50} of 80 nM. Tumor xenograft studies using MCF-7 cells revealed antitumor effects at 10 mg/kg for 30 days. Computational analysis was also applied to increase the bioavailability and three compounds showed 45-60% tumor growth inhibition in vivo compared to the 26% growth inhibition of SR13650. SR13668 was the most potent compound and also displayed antitumor effects in other xenograft models. In vitro, SR13668 was shown to inhibit Akt activation by blocking growth-factor stimulated phosphorylation and showed favorable toxicological profiles [53]. This drug is currently in phase 0 trials for the treatment of cancer [54].

Raveendra et al. used pharmacophore modeling in an effort to identify novel HIV-1 integrase (IN, enzyme mediator of the integration of viral cDNA into the host genome) inhibitors. This model was created with the HipHop algorithm within Catalyst and was based on the Quinolone 3-carboxylic acid class of IN inhibitors that show IC_{50} values ranging from 43.5 to 7.2 nM and EC_{50} against HIV-1 replication of 805 to 0.9 nM [55]. The final pharmacophore hypothesis consisted of four features including a negatively ionizable feature, hydrogen-bond

acceptor, and two hydrophobic aromatic features. 362,260 commercially available compounds were screened and 56 selected for in vitro evaluation. 11 of those tested inhibited the IN catalytic activity with an IC_{50} value < 100 μM. Five compounds had an IC_{50} less than 20 μM and the most potent compound inhibited both the 3' processing (IC_{50} 14 μM) as well as strand transfer activities (IC_{50} 5 μM) of IN[56] . Mugnaini et al created a pharmacophore model and screened the ASINEX database for inhibitors of IN. One compound selected for in vitro testing had a novel scaffold and anti-integrase activity with IC_{50} of 164 μM. Further improvement of this compound yielded an analogue with IC_{50} of 12 μM [57].

Noeske, et al [58] used 2D-pharmacophore-based virtual screening to identify novel mGlu1 antagonists. Antagonism of this receptor has been studied in regards to therapeutic potential in neurodegenerative diseases, anxiety, pain, and schizophrenia [59, 60]. Six reference mGlu1 antagonists were used to construct 2D-pharmacophores with the CATS software package [61]. This software assigns all atoms in a compound as either a hydrogen-bond donor, hydrogen-bond acceptor, positively charged, negatively charged, lipophilic, or non-interest atom type. Then, all compounds of a library are compared with the distances between these different atom types in the reference molecule and similarity scores are calculated to rank molecules that most closely fit this 2D-pharmacophore. Screening the Gold Collection of Asinex Ltd yielded six different hit lists (one for each reference molecule). The top hits were collected from all lists as well as hits that appeared in three or more different lists and 23 compounds were tested experimentally for mGlu1 antagonism. Their most potent compound yielded an IC_{50} of 360 nM and was further optimized to a compound with an IC_{50} of 123 nM.

Quantitative Structure Activity Relationship (QSAR)

Quantitative structure-activity relationship (QSAR) models describe the mathematical relation between structural attributes and target response of a set of chemicals [62]. Classical QSAR is known as the Hansch-Fujita approach and involves the correlation of various electronic, hydrophobic, and steric features with biological activity. In the 1960s, Cowin Hansch and others began to establish QSAR models using various molecular descriptors to physical, chemical, and biological properties focused on providing computational estimates for the bioactivity of molecules [63]. In 1964, Free-Wilson developed a mathematical model relating the presence of various chemical substituents to biological activity (each type of chemical group was assigned an activity contribution) and the two methods were later combined to create the Hansch/Free-Wilson method [64, 65]. The general workflow of a QSAR-based drug discovery project is to first collect a group of active and inactive ligands and then create a set of mathematical descriptors that describe the physicochemical and structural properties of those compounds. A model is then generated to identify the relationship between those descriptors and the ligands' experimental activity maximizing the predictive power. Finally, that model is applied to a library of compounds which are defined with the same descriptors. In this way, experimental activities of these compounds can be predicted and ranked. Success of QSAR, therefore, depends not only on the quality of the initial set of active/inactive compounds, but also on the choice of descriptors and the ability to generate the appropriate mathematical relationship. One of the most important considerations regarding this method is the fact that all models generated will be dependent on the sampling space of the initial set of compounds with known activity and their chemical diversity. In other words, divergent scaffolds or functional groups not represented within this "training" set of compounds will not be represented in the final model

and any potential hits within the library to be screened that contain these groups will likely be missed. Therefore, it is most advantageous to cover a wide chemical space within the training set. For a comprehensive guide on performing a QSAR-based virtual screen, please see the review by Zhang [62].

Descriptor Types

Molecular descriptors can be structural as well as physicochemical and, like molecular fingerprints, can be described on multiple levels of increasing complexity. Information described can include properties such as molecular weight, geometry, volume, surface areas, ring content, rotatable bonds, inter-atomic distances, bond distances, atom types, planar and non-planar systems, molecular walk counts, electronegativities, polarizabilities, symmetry, atom distribution, topological charge indices, functional group composition, aromaticity indices, solvation properties, and many others [24, 66-72]. These descriptors are generated through knowledge-based methods, molecular-mechanical, or quantum-mechanical tools [34] and are classified according to the dimensionality of the chemical representation from which they are computed [73]. One-dimensional descriptors encode properties such as molecular weight, refractivity, and solubility [73]. 2D descriptors are commonly computed from topological representations of molecules while 3D descriptors are obtained from the 3D structure of the molecule [73].

Many two-dimension molecular descriptors are based on graph theoretic indices and represent different aspects of molecular structures. The physicochemical meaning of these indices, however, is unclear and incapable of representing some qualities which are inherently three-dimensional (stereochemistry). Three-dimensional molecular descriptors were developed to address some of these issues [74]. Radial distribution functions (RDFs) are the most popular 3D descriptors. RDFs map the probability distribution to find an atom in a spherical volume of

radius r. In its simplest form, the RDF maps the interatomic distances within the entire molecule. Often it is combined with characteristic atom properties in order to fit the requirements of the information to be represented [66]. RDFs not only provide information regarding interatomic distances between atoms and properties, they reflect other information such as bond distances, ring types, and planar versus non-planar molecules. These functions allow estimation of molecular flexibility through the use of a "fuzziness" coefficient that extends the width of all peaks to allow for small changes in interatomic distances. The equation for a property weighted radial distribution function is shown, where f represents the scaling factor, A_i is the atomic weighting property for atom i, A_j is the atomic weighting property for atom j, B is the "fuzziness" coefficient, r_{ij} is the distance between atoms i and j, and N is the number of atoms in the molecule:

$$g(r) = f \sum_{i}^{N-1} \sum_{j>i}^{N} A_i A_j e^{-B(r-r_{ij})^2}$$

CoMFA [72] is another very popular three-dimension QSAR techniques, established over twenty years ago as a standard technique for constructing three-dimensional models in the absence of direct structural data of the target. In this method, molecules are aligned based on their three-dimensional structures on a grid and the values of steric (VDW interactions) and electrostatic potential energies (Coulombic interactions) are calculated at each grid point. Then a multivariable linear regression or partial least squares model is used to predict activity from these features. Comparative Molecular Similarity Indices (CoMSIA) is an important extension to CoMFA. In CoMSIA, the molecular field includes hydrophobic and hydrogen-bonding terms in addition to the steric and coulombic contributions. Similarity indices are calculated instead of interaction energies by comparing each ligand with a common probe and Gaussian-type

functions are used to avoid extreme values [75]. One important limitation to these methods, however, is that their applicability is limited to static structures of similar scaffolds while neglecting the dynamical nature of the ligands [34]. COMFA and CoMSIA translate the pharmacophore hypothesis directly into a QSAR method.

Hristozov, et al. analyzed the performance of different descriptors across a range of benchmarking datasets and found that the performance of a particular descriptor was often dependent on the activity class. It was found that topological autocorrelation usually offers the best dimensionality/performance ratio. The fusion of the ranked lists obtained with RDF codes and 2D descriptor improved results because RDF codes, while giving similar results, covered different parts of the activity spaces under investigation [76]. Increasing the size of training set beyond 100 compounds did not bring a significant improvement in all scenarios.

Statistical Models

Once a set of descriptors has been established for a set of experimentally verified compounds (active and inactive), a statistical model fits these descriptors to their observed behavior. It is this model that can then be applied to a virtual database of molecules to predict which molecules within that library are likely to be as active as or more active than those in the known set. The relationship between QSAR descriptor and biological activity can be modeled as either a linear or non-linear relationship, depending on the complexity of the system as well as the computational resources available. Linear models are most commonly one of three methods: multivariable linear regression analysis (MLR), principal component analysis (PCA), or partial least square analysis (PLS) [34]. MLR is generally the most time consuming method and involves the stepwise addition or removal of descriptors to find the set that can provide the most accurate predictions. This method can require a large number of training compounds as the general rule of thumb is that 4-5 molecules are required for every descriptor used. PCA

increases the efficiency of MLR by extracting information from multiple variables into a smaller number of uncorrelated variables but results are not always straightforward [77, 78]. PCA can be used with a much smaller set of compounds than MLR. PLS combines MLR and PCA and extracts the dependent variable (biological activity) into new components to optimize correlations [79]. PCA or PLS are commonly used for model development with CoMFA [34].

Machine Learning

Neural networks are the most popular non-linear regression models applied to QSAR-based drug discovery [80]. These models are based on a self-learning algorithm where the neural network learns the relationship between descriptors and biological activity through iterative prediction and improvement cycles [34]. A major drawback of neural networks is the fact that they are sensitive to overtraining resulting in excellent performance within the training set but reduced ability to assess novel compounds.

Another machine learning method, the Support Vector Machine (SVM), separates compounds into groups of actives and inactives. It does this by projecting the descriptors of the training compounds onto a multidimensional feature space where a single "hyperplane" is capable of separating the two groups [81]. SVMs have also been applied to separate proteins into druggable and non-druggable classes. The descriptors generated for proteins can be performed with commercial tools such as PROFEAT and ProtParam and include features such as amino acid composition, secondary structure, solvent accessibility and surface properties, as well as descriptors seen with compound analysis such as hydrophobicity, polarity, polarizability, and charge.

QSAR Application in LB-CADD

QSAR has been used to screen for novel therapeutics in the same way both pharmacophore models and fingerprint similarity methods have been applied to virtual libraries.

Casanola-Martin et al used Dragon software to define descriptors for tyrosinase inhibitors. These descriptors include constitutional, topological, BCUT, Galvez, topological charge, 2D autocorrelations, empirical properties and descriptors, and created a model using linear discriminant analysis. In vitro testing revealed their most potent inhibitor with an IC_{50} of 1.72 µM. This presents a more potent inhibition of tyrosinase than the current reference drug L-mimosine (IC50 = 3.68 µM) [82].

Mueller et al used QSAR and artificial neural networks to identify novel positive and negative allosteric modulators of mGlu5. This receptor has been implicated in neurological disorders including anxiety, Parkinson's disease, and schizophrenia [83, 84]. For the identification of positive allosteric modulators (PAMs), they first performed a traditional high throughput screen of approximately 144,000 compounds. This screen yielded a total of 1,356 hits, a hit rate of 0.94%. The dataset from this HTS was then used to develop a QSAR model that could be applied to a virtual screen. To generate the QSAR model, a set of 1,252 different descriptors across 35 categories were calculated using the ADRIANA software package. The descriptors included scalar, 2-dimensional, and 3-dimensional descriptor categories. A statistical model was created with an artificial neural network and the authors iteratively removed the least sensitive descriptors through several rounds in order to create the optimal set. This final set included 276 different descriptors, including scalar descriptors such as molecular weight up to 3D descriptors including the radial distribution function weighted by lone-pair electronegativity and pi electronegativity. A virtual screen was performed against approximately 450,000 commercially available compounds in the ChemBridge database. 824 compounds were tested experimentally for the potentiation of mGlu5 signaling. Of these compounds, 232 were confirmed as potentiators or partial agonists. This hit rate of 28.2% was approximately thirty

times greater than that of the traditional HTS and the virtual screen took approximately one hour to complete once the model had been optimized[85].

In a separate study, Mueller et al [86] used a similar approach to identify negative allosteric modulators for mGlu5. Rodriguez et al had previously performed a traditional HTS screen of 160,000 compounds for allosteric modulators of mGlu5 and found 624 antagonists [87]. QSAR and artificial neural networks were used to generate predictive models trained on the dataset from Rodriguez et al that were then used to virtually screen for novel negative allosteric modulators of mGlu5. The software package ADRIANA was used to generate all descriptors from 35 different categories and iterative rounds of optimization through input sensitivity analysis were performed resulting in a final set of 763 descriptors. The statistical model was used to virtually screen over 700,000 commercially available compounds in the ChemDiv Discovery database. Hits were filtered for drug-like properties and fingerprint techniques were used to remove hits that were highly similar to provide the most variable set of compounds. 749 compounds were tested in vitro and 27 compounds were found to modulate mGlu5 signaling. This hit rate of 3.6% was a significant increase over the .22% hit rate of the traditional HTS screen. The most potent of the compounds showed IC_{50}'s in vitro of 75 and 124 nM and contained a previously unidentified scaffold.

In addition to predicting the behavior of novel compounds within a virtual library, QSAR has been used as a means to improve the enrichment of compound libraries that will be used in traditional high-throughput screening. While many chemical libraries are constructed in a combinatorial manner, it has been reported that combinatorially synthesized libraries do not cover the chemical space of known drugs and natural products and therefore first-in-class drugs with novel scaffolds will be difficult to find using combinatorial synthesis and high-throughput-screening [88]. Additionally, the number of chemical structures with molecular weight under

500 Daltons has been estimated to be 10^{63}, making it necessary to devise strategies in which compounds can be selectively generated to cover the widest area of chemical space possible [89].

QSAR has also been applied to de novo drug design techniques when structural information regarding the target is unknown. Descriptor and model generation is performed and is used to score the de novo generated molecules in place of other structure-based scoring techniques such as docking. Most commonly, these involve evolutionary algorithms where evolved structures are repeatedly modified and their biological activities are estimated using QSAR models. Modifications are achieved by randomly changing a part of the structure. Ligand-based de novo drug design, however, is less practiced than structure-based de novo design due to the inherent challenges in the absence of receptor structure such as difficulty in extracting relevant information from the ligands alone, ensuring that the generated structures are diverse given an often limited supply of reference structures, and ensuring that generated structures are useful in drug discovery. In addition to QSAR methods for scoring generated molecules, simulated receptors and similarity based methods have been applied [90-95]. Feher et al used 5 selective norepinephrine reuptake inhibitors as a training set to generate 2200 molecules using a combination of structural similarity, 2D pharmacophore similarity, and properties to drive the evolution[96]. One of the top scoring compounds was found to be highly active and has been selected as a lead compound in a project at Neurocrine[96].

Golla, et al. applied QSAR-based methods to the design of novel chemical penetration enhancers (CPEs) to be used in transdermal drug delivery [97]. This group used a genetic algorithm to design novel CPE's. In this paradigm, new molecules are generated based on crossover and mutation operations randomly applied to candidates. All generated molecules were scored based on the QSAR model and predicted property values and the highest scoring

molecules were retained for new rounds of evolution. 272 CPE's were used both to generate the QSAR model as well as provide seed molecules for the genetic algorithm. The QSAR model was created using sequential regression analysis and heuristic analysis using CODESSA and contained a final set of 40 descriptors that optimally predicted properties including skin penetration coefficient, logP, melting point, skin sensitization, and irritation.

The top scoring molecules were validated experimentally for permeation and toxicity using Franz Cell with porcine skin and HPLC analysis as well as toxicity effects on human foreskin fibroblasts and porcine abdominal skin. The study resulted in the identification of 18 novel CPE's, 4 of which showed minimal or no toxic effects [97].

Hoeglund used QSAR modeling combined with synthetic optimization in a follow-up to their most potent hit from a 2008 in silico screen for inhibitors of autotaxin. Autotaxin is an autocrine motility factor and has been linked to cancer progression, multiple sclerosis, obesity, diabetes, Alzheimer's Disease , and chronic pain through the production of LPA [98-103]. Analogues of the lead compound were tested and four of the 30 exhibited IC_{50} less than or equal to the lead. The most potent compound showed 3-fold higher affinity for autotaxin than the lead while another compound showed 2-fold higher affinity [104].

Over the past several decades, over 18,000 QSAR models have been reported for a variety of targets with a variety of descriptors. Hansch et al have carefully collected these into a comprehensive database of QSAR models called C-QSAR [105]. This collection has provided not only access to models for novel applications, but allows the analysis of QSAR models to find areas of problems and improvement demands. Kim et al examined the C-QSAR database for outlier patterns (compounds that showed poor prediction when the average prediction for the model was good) and found that over the 47 QSAR models examined, the number of compounds scoring as outliers ranged from 3% to 36% and 26 of the 47 datasets showed 20% or

more compound outliers [106]. They presented several theories as to why QSAR models are so sensitive to the generation of outliers. One possibility came from analysis of the RCSB protein databank where they discovered examples where related analogs were shown to bind in very different poses. Another explanation presented was the confounding variable inherent in many QSAR-based methods, that of protein flexibility. Protein flexibility may allow odd-shaped compounds to bind and exert an effect by conforming to the structural inconsistencies, presenting difficulty in relating the characteristics of an unusual ligand with those of the more common ligands in the training set [106].

Conclusions

CADD is a useful and important tool in the discovery of novel therapeutics. It can be used to reduce the number of compounds necessary to screen in the search for a novel lead thereby reducing the costs and workload associated with a full-scale HTS endeavor. Virtual HTS projects can be performed using computational tools that are capable of screening many more compounds than traditional HTS at a reduced cost. These tools use sophisticated algorithms to predict activity against a target of interest and prioritize future in vitro and in vivo experiments. Over the past twenty years, CADD has proven to be a viable method in the discovery of novel leads and therapeutics.

Ligand-based CADD describes a branch of CADD that uses sets of known active and inactive compounds against a target of interest to predict activities for novel compounds and screen virtual compound libraries. These methods are preferable when information of the target structure is unknown. Several types of ligand-based CADD, ranging in complexity have been applied to drug discovery. The most naïve of these approaches is molecular fingerprints that aim to describe the presence or absence of specific functional groups. Pharmacophore overlaying

involves the superimposition of known active compounds to map the distribution of chemical features that are common amongst all of the actives. vHTS using pharmacophores searches for compounds containing the same distribution of features as those common to the known actives. QSAR applies a wide range of descriptors that can be scalar, 2D, or 3D in nature to numerically represent the information contained within a molecular structure most important for its biochemical behavior. Statistical or machine learning techniques are applied to construct models that can quantitatively predict behavior of compounds from their numerical representations.

These methods have been applied extensively to vHTS projects, resulting in the discovery of novel, highly potent compounds. Additionally, they have been used to improve the properties of previously identified lead compounds, and aid in hit list prioritization through similarity and clustering analysis. However, LB-CADD continues to see improvements by way of more sophisticated alignment and scoring algorithms, more informative descriptors, and improved model generation. In the following chapter, I present a novel QSAR descriptor that addresses some shortcomings regarding enantioselectivity with traditional 3D-QSAR in an effort to produce models with increased predictability and improved performance in vHTS.

CHAPTER II

BCL::EMAS – Enantioselective Molecular Asymmetry Descriptor for 3D-QSAR

Introduction

Stereoisomers are defined as different molecular species of equal constitution which are separated by energy barriers[107]. For organic molecules, stereochemistry is most frequently caused by carbon atoms with four different substituents. However, other stereo-centers exist such as positively charged nitrogen atoms with four different substituents, double bonds with different substituents on each of the two carbon atoms, stereoisomeric allenes, atropisomeric biphenyls, etc. Enantiomers are a subset of stereoisomers that are defined as non-superimposable mirror images (enantios being Greek for opposite and meros for part). Despite their structural similarities, enantiomers can display very different pharmacological profiles. Stereoisomers that are not enantiomers are called diastereomers. Stereoselectivity is widely prevalent in nature as most proteins are formed from the genetically encoded L-amino acids making small molecule binding pockets enantio-selective[108]. In drug discovery, there are examples in which different enantiomers show different efficacies, e.g. dexrabeprazole[109] and beta blockers[110], and different toxicities, e.g. levobupivacaine[111]. In 1992, the FDA issued a statement requiring that the development of any racemate (mixture of a compound's stereoisomers) carries a justification for the inclusion of both isomers[112] and in the year 2000, chiral drugs accounted for over $100 billion in sales[113]. Between 1985 and 2004, the number of single enantiomer drugs as a percentage of chiral molecules increased from 31.6% to 89.8%[114].

Given the importance of stereoselectivity in drug design, it is necessary that any computational approach to drug discovery distinguishes between stereoisomers. In Structure-Based Computer-Aided Drug Discovery (SB-CADD) stereochemistry is explicitly accounted for as the molecule is docked into a structural model of the protein binding site. The 3D structure of the molecule in complex with the protein is evaluated taking its stereochemistry into account. In complex with the target protein even enantiomers turn into diastereomers and can be distinguished. In Ligand-Based Computer-Aided Drug Discovery (LB-CADD) the chemical structure of active compounds is compared to derive common features that determine activity. The task of distinguishing stereoisomers and in particular enantiomers becomes more challenging as stereochemistry needs to be defined in the absence of the protein. This is impossible in 2D molecular descriptors where only the constitution of a molecule is taken into account. Therefore, extensions to 2D molecular descriptors have been developed – sometimes described as 2.5D descriptors – that describe configuration and can therefore define stereochemistry. Lastly, 3D descriptors based on the molecular conformation can define stereochemistry, if appropriately designed.

The IUPAC convention for distinguishing stereoisomers is the Cahn-Ingold-Prelog (CIP) convention distinguishing R (rectus) and S (sinister) configuration of stereocenters. It requires a priority weighting system for the different substituents that is incapable of dealing with some complex scenarios. Extensions to the CIP system have been introduced to handle situations in which the chiral center did not rest on an atom but rather a chirality plane or axis and for stereoisomers which do not possess centers of chirality at all (stereisomeric allenes, atropisomeric biphenyls, and ansa-compounds)[107]. Further complications arise for pseudoasymmetric stereogenic units, defined as pairs of enantiomorphic ligands together with two ligands which are non-enantiomorphic. In cases such as these, the priorities of two

substituents depend on their own chiral centers. One particular disadvantage is that the CIP nomenclature does not always follow chemical intuition. For example, take the two molecules HC(CH3)(OH)F and HC(CH3)(SH)F. Naively we would align these close derivatives by superimposing H with H, CH3 with CH3, OH with SH and F with F. This assigns R-HC(CH3)(OH)F to S-HC(CH3)(SH)F and vice versa. In result, closely related derivatives that place similar functional groups in the same regions of space and are likely to have similar activity can have opposite CIP assignment. Therefore, the CIP convention is not suitable to describe stereochemistry effectively for LB-CADD.

Extensions to 2D-QSAR have been proposed to distinguish enantiomers. Golbraikh and co-workers introduced a series of chirality descriptors that use an additional term called the chirality correction added to the vertex degrees of asymmetric atoms in a molecular graph [115]. This method is similar to one proposed by Yang and Zhong[116] where the chiral index was instead appended to the substituents attached to the chiral center. Multiple similar algorithms have also been proposed [117-120]. For example, Brown, et al[117] added chirality to their graph kernel method. The drawbacks of these methods include their reliance on the problematic R/S designations as well as the combination of spatial and atom property information such that their indices become a principally mathematical concept with little interpretation on physical terms.

Another approach proposed by Benigni and co-workers [121] describes a chirality measure based on the comparison of the 3D structure for a molecule with all others in a data set. Zabrodsky[122] proposed a similar continuous symmetry measure which quantifies the minimal distance movement for points of an object in order to transform it into a shape of desired symmetry. However, these molecular similarity indices are very sensitive to relative

orientation and depend on pairwise molecular indices which can complicate QSAR-based high throughput screening.

Aires-de-Sousa, et al [123-125] introduced a 3D-QSAR method for handling enantiomers. Classical 3D-QSAR descriptors such as radial distribution functions are incapable of distinguishing between enantiomers based on their nature. This method employs an RDF-like function that utilizes a ranking system for each chiral center introduced by Zhang and Aires-de-Sousa that reinterpreted the CIP rules in terms of more meaningful physicochemical properties. Additionally, it had the benefit of being a vector rather than single value which was equal and opposite for enantiomer pairs. However, this method requires the identification and appropriate labeling of all stereogenic units and suffers from the fact that spatial information is combined with atom properties where some physical interpretability is lost. It is also worth mentioning that it is not clear if it is pharmacologically relevant to specify every stereogenic component of a molecule, but rather if different profiles between enantiomers depend on specific chiral centers and/or an overall chirality of the molecule as a whole.

CoMFA[72] is an appealing method for distinguishing between enantiomers as it avoids the necessity to identify stereogenic centers. Rather, it intrinsically takes chirality into account as the molecular fields of chiral isomers are inherently different. However, the method relies on superimposition of all molecules[115] which is difficult to achieve for large or diverse substance libraries.

Here we propose a novel enantio-selective 3D descriptor for QSAR that is similar to the RDF-like function proposed by Aires-de-Sousa and co-workers but with important differences to address the concerns raised above. We call this new method EMAS (Enantio-selective Molecular ASymmetry). Our method does not rely on any priority ranking or distinction of every stereogenic unit, thereby eliminating the need to combine spatial and atomic properties and

bypassing the difficulties that arise in non-conventional chiral centers. Rather, the enantiomeric distinctions “emerge” from the spatial distribution of atoms within the molecule. Additionally, EMAS is designed to avoid a rigid distinction between enantiomers but rather to represent the overall asymmetry of a molecule as it compares to other similar molecules as well as its own enantiomorphs. Therefore, EMAS intends to describe overall molecular asymmetry while including a directionality component that can distinguish between enantiomers.

Results and Discussion

Shape and Property Enantiomorphism

Enantiomorphism in small molecules is impacted by two phenomena. The first factor is the shape of the molecule – i.e. the distributions of its atom coordinates in space. If the mirror image of this shape cannot be superimposed with the original version, the two molecules are enantiomers. Beyond the overall shape the distribution of properties plays a role. We can envision molecules that have a (near) perfect symmetric shape. Image and mirror image will be identical shape wise. However, distribution of partial charge, polarizability, and electronegativity can be enantiomorph. While both contributions are coupled they represent two dimensions of one phenomenon. For a specific molecule one of the other factors might be more pronounced. For example steroids can have enantiomorph shapes but have relatively uniform property distributions as they are dominated by apolar CH groups. On the other hand, the molecule CFClBrI is an almost perfect regular tetrahedron with a highly enantiomorph distribution of partial charge and polarizability. As both contributions can determine properties and activities of small molecules, stereochemical descriptors should capture and ideally distinguish both contributions.

Radial Distribution Functions separate shape information and property distribution

Radial Distribution Functions (RDFs) are often applied in 3D-QSAR[66, 126]. As a means of comparison, the general form of the atomic radial distribution function is shown:

$$f(r) = \sum_{i}^{n} \sum_{j}^{n-1} P_i P_j e^{-\beta(r-r_{ij})^2}$$

In this equation, β is a smoothing parameter, often called the 'temperature' while r_{ij} is the distance between atoms i and j, n is the total number of atoms in the molecule, and r is the running variable for the function $f(r)$. Often, such equations are 'weighted' with a property coefficient for both atoms $P_i P_j$. The function plots shape (i.e. distance between two atoms) on the x-axis and the respective property coefficient on the y-axis thereby separating geometry from property distribution. With $P_i P_j = 1$ this function is a representation of the overall shape of the molecule based on the frequencies of all atom pair distances within each radial distance step. As distances are invariant to mirroring, enantiomers share identical RDF functions. Note that diastereomers have distinct RDFs as not all atom pair distances are identical.

Expanding RDFs to 'signed' volumes that are sensitive to shape enantiomorphy

We first look for the simplest geometric form that would be sensitive to mirroring. This shape would be a tetrahedron. We choose tetrahedrons consisting of all combinations of three atoms i, j, k and the center of the molecule. Other approaches use all permutations of four atoms. The present approach reduces the computational demand. The geometric property plotted for the tetrahedron is volume. c_i, c_j, and c_k are the coordinates of the three atoms. The center of the molecule is defined by point o. Therefore, the signed volume is computed as:

$$signed\ volume = \frac{1}{6}\left(\overrightarrow{c_i c_j} \times \overrightarrow{c_i c_k}\right) \cdot \overrightarrow{o c_i}$$

While the absolute term always reflects volume it is important to note that the result can have a positive or negative sign, depending on the order of points which is initially arbitrary. We note that the volume has an arbitrary sign that inverts when the molecule is converted into its mirror image. We note further that the volume becomes 0 if the plane defined by c_i, c_j, and c_k includes o. This property is beneficial as a planar arrangement of atoms cannot be enantiomorph.

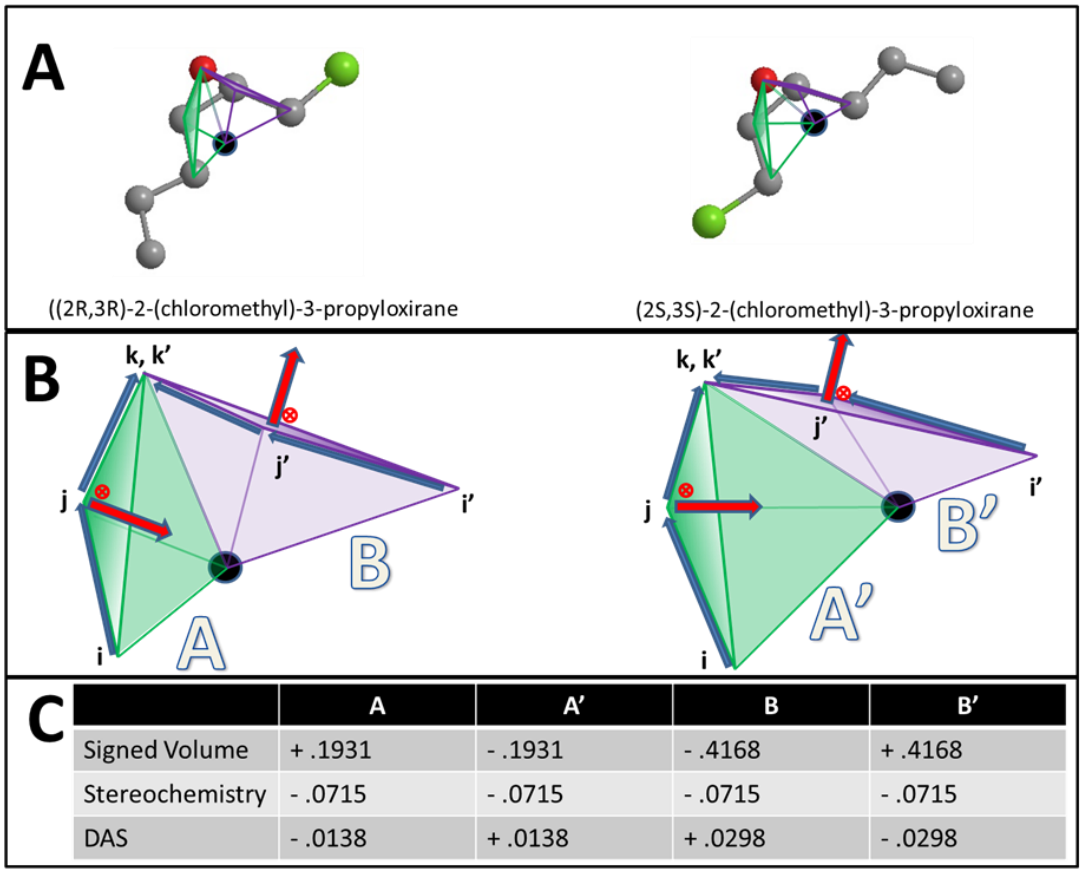
However, for a tetrahedron to contribute to enantiomorphy, its edges $\left\|\overrightarrow{c_i c_j}\right\|$, $\|\overrightarrow{c_i c_k}\|$, and $\left\|\overrightarrow{c_j c_k}\right\|$ must be of different length. This property is captured by a stereochemistry score:

$$stereochemistry = \frac{\left(\left\|\overrightarrow{c_i c_j}\right\| - \|\overrightarrow{c_i c_k}\|\right) * \left(\|\overrightarrow{c_i c_k}\| - \left\|\overrightarrow{c_j c_k}\right\|\right) * \left(\left\|\overrightarrow{c_j c_k}\right\| - \left\|\overrightarrow{c_i c_j}\right\|\right)}{0.0962243 * \max\left(\left\|\overrightarrow{c_i c_j}\right\|, \|\overrightarrow{c_i c_k}\|, \left\|\overrightarrow{c_j c_k}\right\|\right)^3}$$

Two things emerge from the numerator. The asymmetry is evaluated based on the variation in distances between the three atoms. If any two distances are equal, the triangle formed from the three atom coordinates will contain perfect symmetry and the score will be 0. Additionally, the directional (enantiomorphic) information emerges based on the order of distances. For example, if $\left\|\overrightarrow{c_i c_j}\right\| > \|\overrightarrow{c_i c_k}\| > \left\|\overrightarrow{c_j c_k}\right\|$, then this product will have a negative sign $(+) * (+) * (-)$. However, if, from the vantage point of the molecular center, the order of distances has been shuffled (as would be seen in an enantiomer $\|\overrightarrow{c_i c_k}\| > \left\|\overrightarrow{c_i c_j}\right\| > \left\|\overrightarrow{c_j c_k}\right\|$), the sign changes as well $(-) * (+) * (-)$. Recall that by allowing a signed volume, we ensure that the order of distances does not rely on the order of atoms coordinates encountered, but rather as the order of distances seen from the molecular center in terms of the cross-product's direction.

The final directional asymmetry score (DAS) of any given atom triplet becomes:

$$DAS = \sqrt[3]{signed\ volume_{ijk} * stereochemistry_{ijk}}$$



C

	A	A'	B	B'
Signed Volume	+ .1931	- .1931	- .4168	+ .4168
Stereochemistry	- .0715	- .0715	- .0715	- .0715
DAS	- .0138	+ .0138	+ .0298	- .0298

Figure 1. Calculating DAS (**a**) Scores reflect opposing enantiomorphs based on cross-product direction and geometric center. Enantiomers ((2R,3R)-2-(chloromethyl)-3-propyloxirane and (2S,3S)-2-(chloromethyl)-3-propyloxirane) with two stereocenters are shown. (**b**) Two triangles are visualized in both enantiomers. These triangles encompass the same triplets of atoms between the two molecules. Four tetramers formed by the atom triplets and molecular center are visualized. i, j, k, and i', j', k' reflect the order of these atoms in either molecule. Importance of atom ordering is shown based on the direction of cross-product (red arrow) and location of geometric center (black circle). (**c**) Volume and score calculations for four tetrahedrals across both enantiomers are shown. Note the opposite signs and scores between the two enantiomers' tetrahedrals.

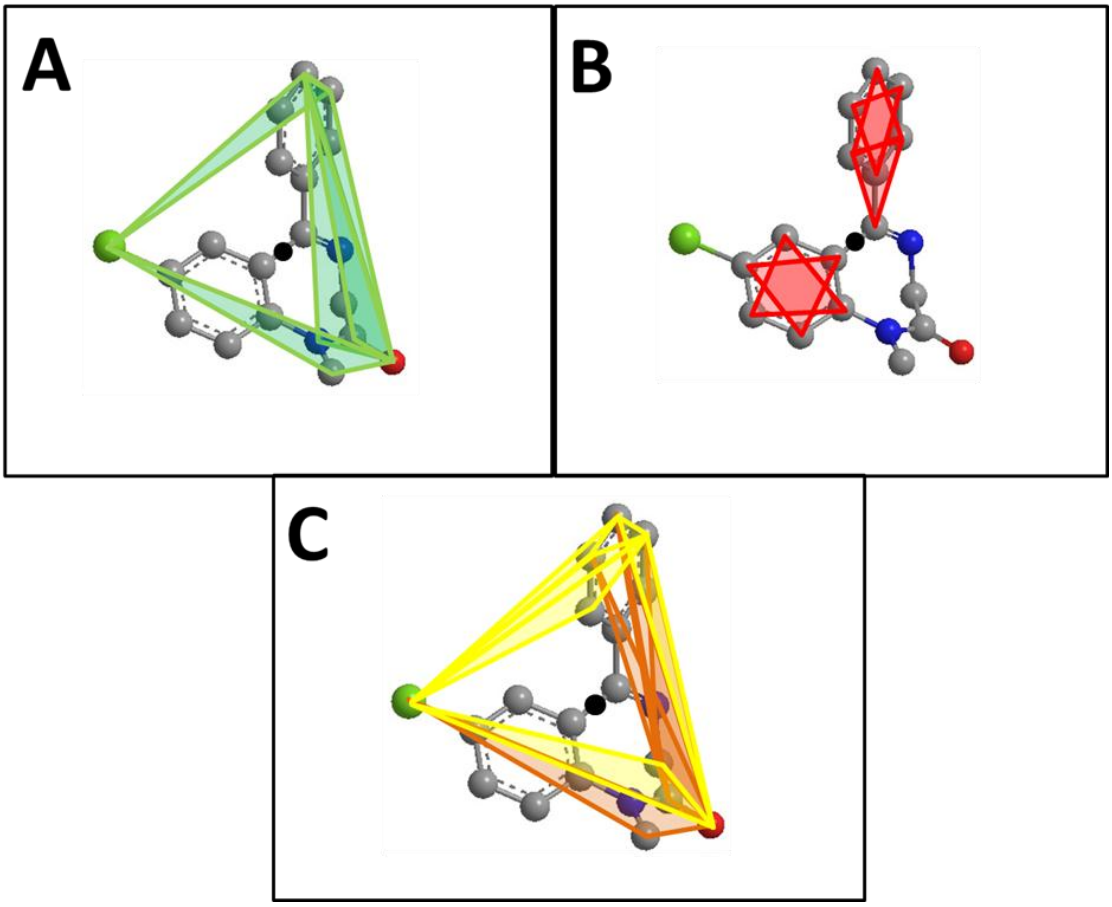


Figure 2. Atom triplets in Diazepam (**a**) Top 5 scoring atom triplets in diazepam are shown. The black spot in all figures represent the geometric center of the molecule. (**b**) Lowest 5 scoring atom triplets in diazepam. All triplets shown here score 0 and do not contribute to the RDF-like code. (**c**) Top 5 positive and top 5 negative scoring triplets in diazepam. Here is visualized the different distribution of high scoring positive (yellow) versus high scoring negative (orange) triplets in diazepam.

Note that for the final DAS, the product's cube-root has been taken to achieve a dimension of distance resembling a common RDF. This procedure preserves the sign and expands the range of frequently occurring low-scoring triplets at the cost of rare triplets with high scores. Substituting this directional asymmetry in place of atom distance, the EMAS function becomes:

$$EMAS(r) = \sum_{i}^{n} \sum_{j}^{n-1} \sum_{k}^{n-2} sign(DAS) \times e^{-\beta\left(r - |DAS_{ijk}|\right)^2}$$

Where β is the smoothing parameter, n is the total number of non-hydrogen atoms, and r is the running variable of the function $EMAS(r)$. The alternate sign preceding the exponential function transfers the "directionality" of the score to the overall function so that at any given score, the intensity reflects the subtraction of negative (one direction) from positive (opposite direction). Figure 3 maps the EMAS plot for epothilone B and its mirror image.

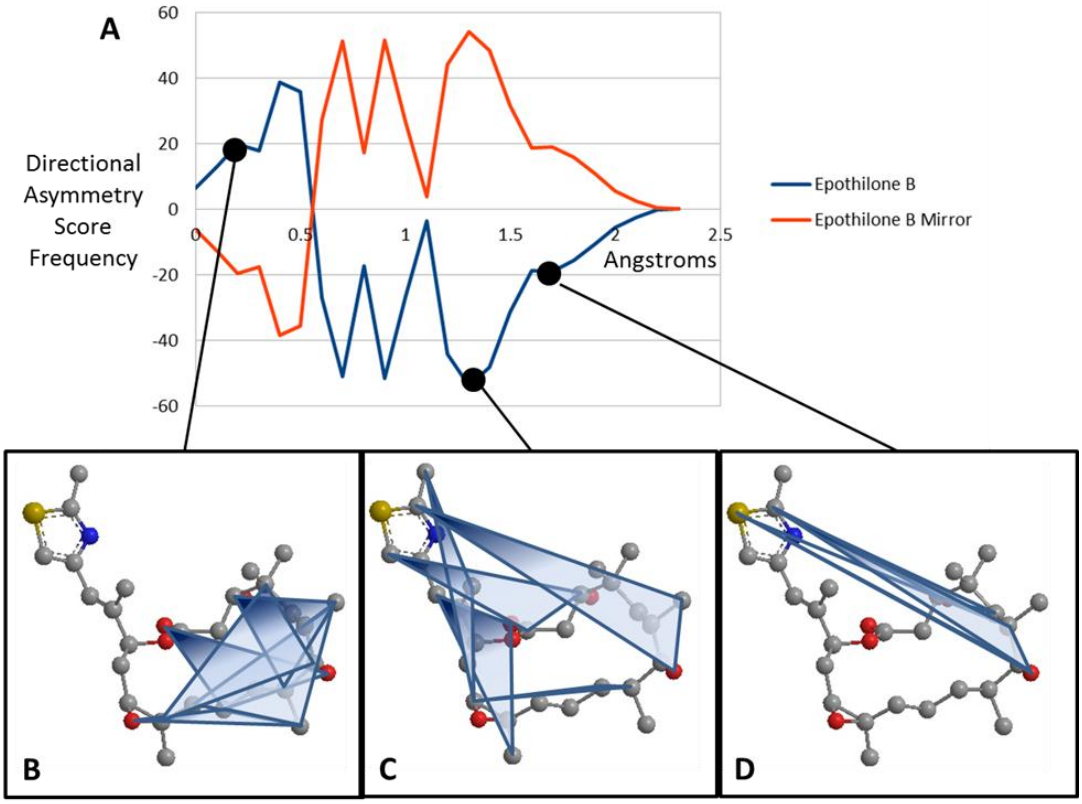


Figure 3. EMAS curves for Epothilone B (**a**) Plotted EMAS curves for Epothilone B (blue) compared with its mirror image (red). X-axis represents the Directional Asymmetry Score in angstroms while the y-axis indicates the frequency of these scores across the entire molecule. (**b**) Atom triplets with a directional asymmetry score of approximately 0.3 angstroms. Note that these triangles generally cover the center of the molecule and are fairly symmetric. (**c**) Atom triplets with a directional asymmetry score of approximately 1.3 angstroms. Note that these triangles are further from the center of the molecule and have an asymmetric shape. **(d)** Atom triplets with a directional asymmetry score of approximately 1.7 angstroms. Note that these atom triplets lie furthest from the center of the molecule and are very asymmetric.

As with the basic radial distribution function, the absence of any weighting coefficient results in a descriptor that encodes only spatial information. While this is important information in and of itself, the addition of a property weighting coefficient increases the utility of this descriptor. Since we are iterating over all atom triplets, the possibility that one atom property can throw off two other atom properties in unintended ways made it problematic in some cases to simply multiply the three atom properties together. Adding the properties, on the other hand, can circumvent this issue but two atom properties of equal magnitude and opposite signs can cancel each other out. Therefore, we retained the functionality for both property coefficient methods and suggest that any use of this descriptor in larger datasets test either method since one may outperform the other depending on the dataset.

The single biggest drawback to our method is that it is highly sensitive to molecular flexibility. However, this is a common drawback seen with many 3D-QSAR techniques. By limiting an algorithm to a single static conformation, compounds that interact with binding sites while in a conformation that is different than the one used can be missed. One possible solution to this issue is through the use of conformational ensembles. In this case, compounds can be represented by a range of conformations rather than a single static conformation. Strategies incorporating conformational ensembles are currently being pursued in our laboratory and are hypothesized to increase the predictability of this descriptor, especially in molecules with a high degree of flexibility.

Evaluation of EMAS as a Novel Descriptor

Predictability Benchmarking: Cramer's Steroids

A commonly used dataset for evaluating the predictive capability of novel stereochemistry-based descriptors was introduced by Cramer et al. in 1988 [127] and several

structures were corrected in a subsequent publication [128]. These thirty-one steroid structures are accompanied with their experimental binding affinities to human corticosteroid-binding globulins (CGB) and provide a small dataset containing many stereocenters. Additionally, the rigidity of these compounds makes them an ideal benchmark set for 3D-QSAR algorithms eliminating the factor of conformational flexibility. Since EMAS can be employed in three forms: spatial only, property weighting coefficient via summation, and property weighting coefficient via multiplication, we trained three separate artificial neural network (ANN) models using descriptors derived in each of these three methods. To predict binding affinities over the entire dataset, we used a cross-validated leave-one-out approach. To compare the predictive power of our model versus other descriptors that have been tested against the steroid set, we calculated the correlation coefficient r^2 of predicted versus actual affinities and the "cross-validated r^2" q^2.

As expected, the ANN model generated using no property weighting (solely spatial information) performed the worst of the three, producing a r^2 of 0.78 and a q^2 of 0.60. By weighting with a multiplicative property coefficient, the performance increased considerably, resulting in a r^2 of 0.86 and a q^2 of 0.74. Weighting with the property summation coefficient, yielded the best predictions with a r^2 of 0.89 and a q^2 of 0.78.

Since we began with an interest in generating a molecular asymmetry descriptor that could distinguish between enantiomers, we wanted to ensure that the inclusion of directionality increased the information contained in the descriptor. Therefore, we created a version of the descriptor that incorporates just the absolute value of all stereochemistry scores, thereby eliminating all directional information while retaining all other spatial information. We found that by training our model without directional information, the predictive capabilities for the steroid affinities decreased to a r^2 of 0.65 and a q^2 of 0.41, reinforcing our original design to

capture stereochemistry. We also compared the model employing EMAS with one created with a traditional RDF. This model performed worse than any of our three methods giving a r^2 of 0.75 and a q^2 of 0.56. Weighting the RDF's with the same properties used to weight EMAS did not produce any significant improvement in the model (data not shown). Cross-validated predictions for all variations of EMAS as well as the experimental affinities can be found in Table 1.

Table 1. Experimental and predicted binding affinities for the 31 Cramer's steroids using novel stereoselective descriptor to train ANN models. Spatial predictions utilize the novel descriptor without any atom property weighting. Multiply properties utilize the novel descriptor weighted by the product of atom properties. Sum properties utilize the novel descriptor weighted by the sum of atom properties.

Molecule	Observed CBG affinity (pKa)	Predicted [spatial]	Predicted [multiply properties]	Predicted [sum properties]	Predicted [no stereo-chemistry]
aldosterone	-6.28	-7.47	-7.31	-7.25	-7.22
androstanediol	-5.00	-5.47	-5.46	-5.33	-5.56
5-androstenediol	-5.00	-5.47	-5.43	-5.36	-5.75
4-androstenedione	-5.76	-5.64	-5.60	-5.79	-6.36
androsterone	-5.61	-5.78	-5.81	-5.55	-5.42
corticosterone	-7.88	-7.30	-7.37	-7.32	-7.34
cortisol	-7.88	-7.63	-7.58	-7.64	-7.33
cortisone	-6.89	-7.22	-6.83	-7.39	-7.07
dehydroepiandrosterone	-5.00	-5.39	-5.13	-5.46	-5.80
11-deoxycorticosterone	-7.65	-7.48	-7.47	-7.50	-6.85
11-deoxycortisol	-7.88	-7.66	-7.53	-7.59	-7.52
dihydrotestosterone	-5.92	-5.38	-5.70	-5.43	-5.96
estradiol	-5.00	-5.40	-5.36	-5.32	-5.21
Estriol	-5.00	-5.25	-5.26	-5.43	-6.10
estrone	-5.00	-5.30	-5.21	-5.54	-5.42
etiocholanolone	-5.23	-6.42	-6.44	-6.22	-6.27
pregnenolone	-5.23	-5.30	-5.25	-5.37	-6.37
17a-hydroxypregnenolone	-5.00	-5.20	-5.28	-5.29	-6.65
progesterone	-7.38	-7.17	-7.27	-7.13	-6.46
17a-hydroxyprogesterone	-7.74	-7.42	-7.39	-6.97	-6.70
testosterone	-6.72	-6.08	-6.36	-6.19	-5.94
prednisolone	-7.51	-7.61	-7.36	-7.65	-7.03
cortisolacetat	-7.55	-6.74	-6.90	-7.63	-6.00
4-pregnene-3,11,20-trione	-6.78	-6.40	-6.83	-6.09	-6.46
epicorticosterone	-7.20	-5.98	-6.00	-7.03	-7.15
19-nortestosterone	-6.14	-5.58	-5.86	-5.54	-5.45
16a,17a-dihydroxyprogesterone	-6.25	-7.25	-7.04	-7.46	-7.36
16a-methylprogesterone	-7.12	-6.69	-6.39	-6.78	-6.60
19-norprogesterone	-6.82	-6.01	-6.30	-7.25	-6.19
2a-methylcortisol	-7.69	-6.62	-7.22	-7.68	-6.57
2a-methyl-9a-fluoro-cortisol	-5.80	-7.56	-6.97	-6.22	-6.74
	r^2	0.78	0.86	0.89	0.65
	q^2	0.60	0.74	0.78	0.42

Since this dataset is well-established across similar descriptors in the literature, we compared our predictive power to other methods and found that our best q^2 fell at the average q^2 of all of these methods ($0.63 < q^2 < 0.94$). This result is somewhat difficult to interpret for several reasons: a) different statistical models are utilized, b) different degrees of cross-validation were employed, and c) our descriptor solely describes stereochemistry and is meant to be complemented by other descriptors (read below). Most of the competing descriptors include more information on molecule size, shape, and property distribution. However, it is important to note that while EMAS does not require any molecular alignment or pre-annotated stereocenters, it is capable of performing well with a dataset that contains a great deal of stereochemistry. Additionally, the inclusion of directional information outperforms a similar implementation lacking directional information as well as the similar RDF descriptor weighted with or without atom properties. For a comparison of our $q^2$with other documented tests against Cramer's steroids, see table 2.

Table 2. Comparison of novel stereoselective descriptor predictability with other published QSAR methods against the Cramer's steroid set. Calculation of q^2 can be found in the methods section. Statistical model generation method is indicated as well as QSAR method employed are indicated for each reference.

QSAR Method	Model Creation	q^2	Reference
Purely Spatial RDF-like stereochemistry	Artificial Neural Network	**0.56**	
Property weight RDF-like stereochemistry (product)	Artificial Neural Network	**0.74**	
Property weight RDF-like stereochemistry (sum)	Artificial Neural Network	**0.78**	
Stochastic 3D-chiral linear indices	Multiple Linear Regression	0.87	[119]
Chiral Topological Indices	Stepwise Regression Analysis	0.85	[116]
Chiral Graph Kernels	Support Vector Machine	0.78	[117]
Chirality Correction and Topological Descriptors	K-nearest neighbor	0.83	[115]
Molecular Quantum Similarity Measures	Multilinear Regression	0.84	[129]
Shape and Electrostatic Similarity Matrixes	Non-linear Neural Network	0.94	[130]
Comparative Molecular Moment Analysis	Partial Least Squares (PLS)	0.83	[128]
Comparative Molecular Similarity Indices Analysis	PLS	0.67	[131]
Comparative Molecular Field Analysis	PLS	0.65	[127]
E-state Descriptors	PLS	0.62	[132]
Molecular Electronegativity Distance Vector	Genetic Algorithm PLS	0.78	[133]
Molecular Quantum Similarity Measures	Multilinear Regression and PLS	0.80	[134]

vHTS Utility and Enrichment Benchmarking: PUBMED AID891

We provide the above analysis for comparison. However, realistically the steroid dataset is too small to provide a good benchmark for EMAS as often the number of features (24) is in the same order of magnitude as the number of data points (31). Therefore we tested the descriptor in a more virtual high-throughput screening (vHTS) endeavor. For the benchmark dataset, we used publicly available results of a conformational screen for inhibitors and substrates of cytochrome P450 2D6 (AID 891). This dataset is of moderate size (approximately

10,000 molecules) and contains both active (18%) and inactive (82%) compounds. We employed a forward-feature selection (FFS) analysis that selects optimal descriptors from RDF's, 3D Autocorrelations (3DA), and 2D Autocorrelations (2DA) labeled with atom properties including charge, electronegativity, and effective polarizability (see Experimental Section). For a complete list of features tested in forward-feature selections, please see appendix table A1. ANN 3D-QSAR models were trained with and without inclusion of the EMAS descriptors in the list of descriptors for FFS to choose from. Hence the utility of the EMAS descriptor can be evaluated in two ways: a) are the EMAS descriptors selected by the FFS procedure? and b) does the final model that includes EMAS descriptors have increased predictive power? The FFS with the default set of initial features resulted in a best descriptor set of 9 features distributed evenly across RDF's, 3D Autocorrelations (3DA), and 2D Autocorrelations (2DA). Cross-validated predictions from the ANN model constructed with this feature set produced an enrichment of 3.94 and a receiver operating characteristic (ROC) curve with an area under the curve (AUC) of 0.826.

An identical FFS analysis was performed by combining the default set of features with 34 EMAS features including all three variations of EMAS (spatial, property weighting via sum, and property weighting via product) weighted with the same list of properties used to test RDFs, 3DAs, and 2DAs. The best set of features contained 20 total features distributed across RDF's, 3DA's, 2DA's, number of hydrogen bond donors, and several EMAS features. There were a total of seven EMAS features represented in the best feature set. Therefore, almost one third of the total features in the best feature set generated through this analysis were EMAS features. This set of seven features contained a spatial EMAS weighted by Van der Waals surface areas, three EMAS features weighted via the product method and three EMAS features weighted via the sum method. This substantial representation of EMAS in the best feature set suggests that EMAS

successfully provides useful information for the model development that may not be represented in any other feature in the original set. Cross-validated predictions from the ANN model constructed from this EMAS-inclusive feature set produced an enrichment of 4.38 and a ROC curve with an area under the curve of 0.837, a clear improvement over the control model. Positive predictive value (PPV) is a related measure of a model's predictive capability which tracks predictive precision as more and more positive predictions are made. By comparing the average PPV precision over a range of the fraction of total predictions made (fraction positive predictions, FPP) of interest, it is possible to compare predictive capabilities for two models. Over the FPP range of .005 to .05, we find that our model trained with the EMAS features performed significantly better than the model trained without EMAS features (.727 PPV precision compared with .651). A paired t-test for the cross-validated models comparing precisions in this FPP range showed that this is a statistically significant improvement ($p < .005$) over the analysis completed without EMAS features. For a complete list of the best features determined from both forward feature analyses, please see the appendix table A2. Comparative ROC and PPV curves from the forward feature analyses for the control set of features and the control set combined with EMAS features are shown in figure 4.

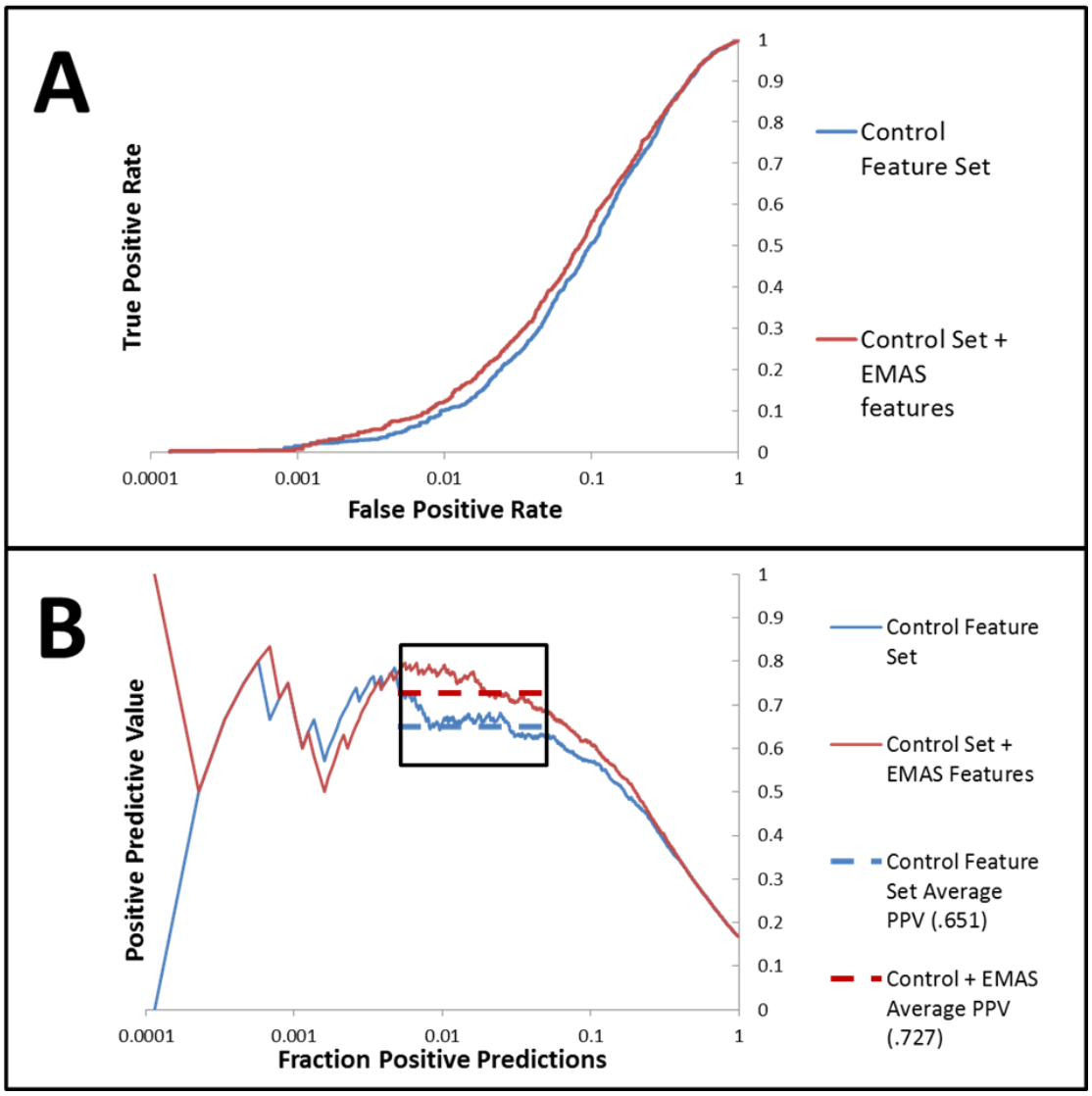


Figure 4. ROC and PPV results for the feature forward analysis with the control set of features compared with the control set combined with EMAS features (**a**) AID891 prediction ROC curves generated from the ANN models trained with the best descriptor set generated from the forward feature analysis beginning with the control set of features combined with the novel EMAS features (red) show improved performance when compared with ROC curves generated from the ANN models trained with the best descriptor set generated from the forward feature analysis beginning with the control set of features (blue) (**b**) PPV curves for models trained with the best descriptor set of control features combined with the EMAS features (red) shows improved performance over those models trained with the best descriptor set of control features only (blue). Dashed lines of corresponding colors show the average PPV values over the FPP region from which the models were optimized (.005 to .05 fraction positive predicted values).

Conclusions

The goal of this project was to develop a 3D-QSAR descriptor that was capable of not only distinguishing between enantiomers but of describing the overall degree of asymmetry for a molecule. This was accomplished by developing an RDF-like curve that described the distribution of 'directional asymmetry scores (DAS)' rather than inter-atomic distances. The DAS is designed to incorporate information regarding the degree and direction of asymmetry between each atom triplet in the molecule. The degree of asymmetry is calculated as a product of how asymmetrically the three atoms are distributed and the distance they lie from the center of the molecule. This asymmetry is related to the differences between their interatomic distances and the distance from the center of the molecule is related to the volume of the tetrahedron created by the three atom coordinates and the geometric center of the molecule. The direction of asymmetry is related to the distribution of the interatomic distances between these three atom coordinates from the point of view of the center of the molecule. If the sides of the triangle created by these three atoms are different, then identical triangles "pointing" in opposite directions will have a different ordering of sides depending on which direction they "point." This is the key variable that allows the descriptor to distinguish between enantiomers. To exclude any influence that the order in which atoms are listed in the molecule may play on this directionality scheme, we offset this by incorporating the cross-product of the two vectors created from the three atoms. This cross-product will swap signs were the atoms are ordered differently thereby eliminating the influence of the order of atoms.

We tested the value of this descriptor by training ANN 3D-QSAR models. In order to provide a basis of comparison with other documented QSAR methods that address stereoselectivity, we used a small dataset of steroids that is commonly used as a benchmark for these types of descriptors. We found that the predictability of our descriptor performed

comparably with other stereochemistry-based descriptors when evaluated with this set of 31 steroids (r^2 = 0.89, q^2 = 0.78). Additionally, we assessed the utility of the EMAS descriptor by running vHTS experiment on a publically available dataset (PUBCHEM AID 891). A forward-feature selection analysis that determines the most effective set of descriptors for this dataset was employed and the best set of features included several EMAS functions (7 EMAS of 20 total features). This set of features improved the performance of our models over those that were tested without EMAS functions (enrichment of 4.38 when including EMAS versus enrichment of 3.94 without EMAS).

We conclude that the EMAS descriptor encodes stereochemistry thereby providing important information that is not captured in other 3D-QSAR descriptors. There are several published QSAR methods that performed better than EMAS for the steroid dataset but these methods often require some heuristic for describing the stereocenters within each of the molecules or aligning the 3D structures of these molecules. By avoiding the necessity to assign a directional designation to each stereocenter, EMAS is capable of evaluating molecules without the problematic R/S annotation method. Other methods that avoid these annotations present their own limitations, as an additional alignment step and framework similarity is required. These issues limit the range of descriptor applicability and introduce more degrees of freedom. While our descriptor is outperformed by multiple techniques with the steroid dataset, we contend that this is not a very accurate comparison. The cross-validation methods used by many of the other methods vary and are often more forgiving than ours. Some drawbacks that EMAS addresses are not encountered with the steroid dataset as it is composed of very similar molecules with relatively simple, fully annotated stereocenters. Additionally, EMAS achieves a more global representation of stereochemistry that retains a physical basis and is applicable to

any set of molecules. This broad applicability is not found with other stereochemistry-based descriptors.

In summary, EMAS provides a widely applicable stereo-sensitive descriptor that intrinsically incorporates physical chirality rather than describing it with abstract annotations.

Methods

Generation of Numerical Descriptors for QSAR Model Creation

3D models of all small molecules were generated using the CORINA software package unless already defined. For feature selection analysis, a set of 2,100 numerical descriptors was generated using the BioChemical Library (BCL) software created in our lab. The descriptors can be classified into 5 categories, including six scalar descriptors (molecular weight, number of hydrogen bond donors, number of hydrogen bond acceptors, logP, total charge, and topological surface area), 18 2-dimensional auto-correlation functions, 18 3-dimensional autocorrelation functions, 18 radial distribution functions, and 34 novel EMAS descriptors. These 34 descriptors included spatially-based asymmetry functions with and without Van der Waals (VDW) surface area scaling, 16 property-weighted asymmetry functions based on the multiplicative scheme, and 16 property-weighted asymmetry functions based on the additive scheme. These properties included sigma charge [135-137], pi charge [138-140], Vcharge [141], total charge [135-140], sigma electronegativity [135-137], pi electronegativity [138-140], effective polarizability [142-144], and lone pair electronegativity [138-140] with and without VDW surface area scaling. The control comparison forward feature selection analysis was performed with a feature set that included all features listed above except the novel EMAS features. This feature set contains 1284 features. For steroid binding predictions, descriptor sets were created using only one

EMAS method and those including property weighting coefficients used the same properties listed.

Training, Monitoring, and Independent Dataset Generation

Cramer's Steroids

The dataset was split for ANN training into three subsets: training, monitoring, and independent. The monitoring dataset is necessary to prevent over-training. Because of the small size of the dataset, only one molecule was labeled independent. Five molecules were used as the monitoring dataset, 25 for training. The set of five molecules was incremented through the entire dataset for a total of 6 different monitoring sets. Leave-one-out cross-validation was performed where each molecule was used as the independent molecule while the remaining 30 molecules were used for training and monitoring. The predictions were averaged across the different monitoring sets to yield the final activity predictions for the entire set of 31 molecules.

PUBMED AID891

AID 891 is a publically available dataset that can be found at http://pubchem.ncbi.nlm.nih.gov/. It contains 1623 active compounds and 7756 inactive compounds tested for inhibition of cytochrome P450 2D6. This dataset was split into 10 clusters, 8 of which were used as the training dataset, 1 used for monitoring, and 1 used for independent. For cross-validation, the monitoring and independent datasets are iterated and then the resulting independent predictions are averaged to give the final list of predicted activities that spans the entire dataset. In order to maximize model performance, the dataset was balanced through oversampling. In other words, the active compounds were represented multiple times so that the number of active compounds roughly equals the number of inactive compounds. This method of balancing has been used to maximize QSAR models in other

datasets where the number of active compounds is significantly less than the number of inactive compounds[86].

The pIC_{50} values of each compound within AID891 and the steroid binding data for the Cramer dataset were used as output for the ANN models. For the AID891 dataset, inactive compounds were set to a pIC_{50} value of 3. The root-mean-square deviation (RMSD) between predicted and experimental activities was used as the objective function for training the ANN.

Artifical Neural Network (ANN) architecture and training

For the AID891 dataset, the ANN was trained using a sigmoid transfer function with a simple weight update of $\eta = 0.1$ and $\alpha = 0.5$. The hidden layer contained eight neurons. For the steroid dataset, the ANN was trained using the same protocol as the AID891 dataset but the number of hidden neurons was reduced to 4 due to the dataset's much smaller size of.

Forward-feature selection for optimal descriptor set selection

Descriptor selection was performed to test the novel descriptor against all other implemented descriptors to see if it provided an increase to enrichment over any of the other descriptors. The approach begins with a single descriptor, trains a model with only that descriptor, and then continuously adds more descriptors one at a time, training a new model each round. At the completion of each round, the descriptor set that produced the lowest RMSD score was retained for the next round. All descriptors not present in the retained list of descriptors are then added individually to that retained list of descriptors and the descriptor set producing the best RMSD score is retained for the next round, and so on. At the completion of these iterations, the round that produced the best RMSD score overall is recalled as the top

descriptor set. If a descriptor appears in this list of best descriptors, then it suggests that significant information had been gleaned from that descriptor during the ANN training.

Model Evaluation

ANN models using the AID891 datasets were analyzed using receiver operation characteristic (ROC) curves to assess their predictive power. These curves plot the rate of true positives versus the rate of false positives as a fraction of the total number of positives. Therefore, a slope of 1 would reflect random guesses as each true positive would be statistically likely to be followed by a false positive. An increase in slope and area under the curve would indicate an increase in predictive power. The initial section of the ROC curve is often most important because it represents compounds with the highest predicted activity. Therefore, enrichment values are determined based on the slope of the ROC curve comprising the first subset of molecules. Increases in enrichment is often the most important measure for application of virtual screening in drug discovery as it reflects the expected factor at which the fraction of actives will be increased over an unbiased dataset.

Positive predictive value (PPV) is a measure related to enrichment which tracks the model's predictive precision as the fraction of predicted positives (FPP) increases from highest predicted activity to lowest. A model is likely to lose precision as the predicted activities approach the cutoff point and therefore it is common to specify a range of FPP of interest when measuring a PPV. FPP is calculated as the number of true positive predictions plus the number of false positive predictions divided by the size of the dataset. PPV is calculated as the number of true positive predictions divided by the total number of positive predictions (true and false positive).

To determine the statistical significance for the average PPV improvement over the FPP range of .005 to .05, we compared the average PPV within this FPP range for each combination of training and modeling datasets that went into the cross-validated model. By aligning these datasets between the two models, we were able to perform a two-tailed paired t-test to show a significant improvement for the cross-validated model including EMAS features over the cross-validated model without EMAS features.

To evaluate the utility of models trained with the steroid dataset in a way which could be comparable with published methods, the conventional correlation coefficient r^2 of the predicted activities against actual activities and cross-validated r^2, also known as q^2 were calculated for each descriptor set. All predicted values used in these analyses were the average predicted activities from each of the leave-one-out models with the different monitoring datasets. The q^2 is calculated from the equation

$$q^2 = \frac{SD - press}{SD}$$

Here, SD is the sum of squared deviations of each biological property from their mean and $press$ (predictive residual sum of squares) is the sum of the squared differences between the actual biological property and the cross-validated predicted property.

Implementation

The descriptor generation and ANN algorithms were implemented in the BioChemistryLibrary (BCL). The training method used is simple propagation, a supervised learning approach. The BCL is an in house developed object oriented Library written in the C++ programming language.

APPENDIX

NORMALIZATION OF STEREOCHEMISTRY SCORE

The stereochemistry score is normalized based on the maximum possible stereochemistry score which can be computed assuming $a \geq b \geq c$ and $c = a - b$

$$f(a,b,c) = -(a-b)(b-c)(c-a)$$

$$= -a^3\left(1-\frac{b}{a}\right)\left(\frac{b}{a}-\frac{c}{a}\right)\left(\frac{c}{a}-1\right)$$

$$= a^3\left(1-\frac{b}{a}\right)\left(2\frac{b}{a}-1\right)\left(\frac{b}{a}\right)$$

With a^3 being a constant and $x := \frac{b}{a}$ we find: $f(x) = 3x^2 - x - 2x^3$.

$$\frac{\partial f}{\partial x} = 6x - 1 - 6x^2$$

$$0 = x^2 - x + \frac{1}{6}$$

$$x = \frac{1 \mp \sqrt[2]{\frac{2}{6}}}{2} \quad \rightarrow \quad x = \frac{1}{2} \mp \sqrt[2]{\frac{1}{12}}$$

$$b = 0.211328, c = 0.788675$$

$$\max\{(1-b)(b-c)(c-1)\} = 0.0962243$$

FORWARD-FEATURE SELECTION DESCRIPTORS

Table A1: Complete list of features used for the forward-feature selection of descriptors. Novel EMAS functions were excluded for the control forward-feature selection.

	Descriptor Name	**Description**
Scalar descriptors	Weight	Molecular weight of compound
	HbondDonor	Number of hydrogen bonding acceptors derived from the sum of nitrogen and oxygen atoms in the molecule
	HBondAcceptor	Number of hydrogen bonding donors derived from the sum of N-H and O-H groups in the molecule
	TopologicalPolarSurfaceArea	Topological polar surface area in [$Å^2$] of the molecule derived from polar 2D fragments
	LogP	Octanol/water Partition coefficient calculated by atom-additive method[145]
	TotalCharge	Sum of atomic formal charges across molecule
Vector descriptors	Identity	weighted by atom identities
2D Autocorrelation (11 descriptors)	SigmaCharge	weighted by σ atom charges
	PiCharge	weighted by π atom charges
3D Autocorrelation (12 descriptors)	TotalCharge	weighted by sum of σ and π charges
	SigmaEN	weighted by σ atom electronegativities
Radial Distribution Function (48 descriptors)	PiEN	weighted by π atom electronegativities
	LonePairEN	weighted by lone pair electronegativities
Novel EMAS Function weighted by sum of properties (24 descriptors)	EffectivePolarizability	weighted by effective atom polarizabilities
Novel EMAS Function weighted by product of properties (24 descriptors)	Vcharge	weighted by partial atomic charges accounting for alternate resonance forms[141]
Every Vector descriptor available with and without van der Waals surface area weighting		

Table A2: Top feature sets following forward-feature selection for both conditions

Control feature selection (without EMAS)		**Novel feature selection (with EMAS)**	
Descriptor Type	**Weight**	**Descriptor Type**	**Weight**
Radial Distribution Function	AtomIdentity [surface area scaled]	Radial Distribution Function	AtomIdentity [surface area scaled]
Radial Distribution Function	Vcharge	Radial Distribution Function	Vcharge
Radial Distribution Function	EffectivePolarizability [surface area scaled]	EMAS (product weight)	AtomIdentity [surface area scaled]
3D Autocorrelation	SigmaCharge	2D Autocorrelation	SigmaEN [surface area scaled]
Radial Distribution Function	LonePairEN	Radial Distribution Function	PiEN [surface area scaled]
2D Autocorrelation	SigmaEN	Scalar	HbondDonor
3D Autocorrelation	SigmaEN	EMAS (product weight)	SigmaEN [surface area scaled]
3D Autocorrelation	Vcharge [surface area scaled]	2D Autocorrelation	EffectivePolarizability [surface area scaled]
2D Autocorrelation	Vcharge [surface area scaled]	3D Autocorrelation	Vcharge [surface area scaled]
		Radial Distribution Function	PiEN
		3D Autocorrelation	SigmaCharge
		2D Autocorrelation	EffectivePolarizability
		EMAS (sum weight)	Vcharge [surface area scaled]
		EMAS (product weight)	Vcharge
		EMAS (sum weight)	TotalCharge
		Radial Distribution Function	EffectivePolarizability
		EMAS (sum weight)	LonePairEN
		EMAS (product weight)	PiEN [surface area scaled]
		3D Autocorrelation	PiEN [surface area scaled]
		Radial Distribution Function	SigmaCharge

REFERENCES

1. Van Drie, J.H., *Computer-aided drug design: the next 20 years.* Journal of computer-aided molecular design, 2007. **21**(10-11): p. 591-601.

2. Doman, T.N., et al., *Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B.* Journal of medicinal chemistry, 2002. **45**(11): p. 2213-21.

3. Vijayakrishnan, R., *Structure-based drug design and modern medicine.* Journal of postgraduate medicine, 2009. **55**(4): p. 301-4.

4. Talele, T.T., S.A. Khedkar, and A.C. Rigby, *Successful applications of computer aided drug discovery: moving drugs from concept to the clinic.* Current topics in medicinal chemistry, 2010. **10**(1): p. 127-41.

5. Hartman, G.D., et al., *Non-peptide fibrinogen receptor antagonists. 1. Discovery and design of exosite inhibitors.* J Med Chem, 1992. **35**(24): p. 4640-2.

6. Sawyer, J.S., et al., *Synthesis and activity of new aryl- and heteroaryl-substituted pyrazole inhibitors of the transforming growth factor-beta type I receptor kinase domain.* Journal of medicinal chemistry, 2003. **46**(19): p. 3953-6.

7. Singh, J., et al., *Successful shape-based virtual screening: the discovery of a potent inhibitor of the type I TGFbeta receptor kinase (TbetaRI).* Bioorganic & medicinal chemistry letters, 2003. **13**(24): p. 4355-9.

8. Shekhar, C., *In silico pharmacology: computer-aided methods could transform drug development.* Chemistry & biology, 2008. **15**(5): p. 413-4.

9. Horvath, D., *A virtual screening approach applied to the search for trypanothione reductase inhibitors.* Journal of Medicinal Chemistry, 1997. **40**(15): p. 2412-2423.

10. Ripphausen, P., et al., *Quo vadis, virtual screening? A comprehensive survey of prospective applications.* Journal of medicinal chemistry, 2010. **53**(24): p. 8461-7.

11. Enyedy, I.J. and W.J. Egan, *Can we use docking and scoring for hit-to-lead optimization?* Journal of computer-aided molecular design, 2008. **22**(3-4): p. 161-8.

12. Jorgensen, W.L., *The many roles of computation in drug discovery.* Science, 2004. **303**(5665): p. 1813-8.

13. Johnson, M.A., G.M. Maggiora, and American Chemical Society. Meeting, *Concepts and applications of molecular similarity*. 1990, New York: Wiley. xix, 393 p.

14. Auer, J. and J. Bajorath, *Molecular similarity concepts and search calculations.* Methods in molecular biology, 2008. **453**: p. 327-47.

15. Willett, P., *Similarity-based virtual screening using 2D fingerprints.* Drug discovery today, 2006. **11**(23-24): p. 1046-53.

16. Hutter, M.C., *Graph-based similarity concepts in virtual screening.* Future medicinal chemistry, 2011. **3**(4): p. 485-501.

17. Bajorath, J., *Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening.* J Chem Inf Comput Sci, 2001. **41**(2): p. 233-45.

18. Bajorath, J., *Integration of virtual and high-throughput screening.* Nat Rev Drug Discov, 2002. **1**(11): p. 882-94.

19. Barnard, J.M. and G.M. Downs, *Chemical fragment generation and clustering software.* Journal of Chemical Information and Computer Sciences, 1997. **37**(1): p. 141-142.

20. Ihlenfeldt, W.D. and J. Gasteiger, *Hash codes for the identification and classification of molecular structure elements.* Journal of Computational Chemistry, 1994. **15**(8): p. 793-813.

21. MDL, MDL Information Systems, Inc.: 14600 Catalina Street, San Leandro, CA, 94577.

22. Durant, J.L., et al., *Reoptimization of MDL keys for use in drug discovery.* J Chem Inf Comput Sci, 2002. **42**(6): p. 1273-80.

23. McGregor, M.J. and P.V. Pallai, *Clustering of large databases of compounds: Using the MDL ''keys'' as structural descriptors.* Journal of Chemical Information and Computer Sciences, 1997. **37**(3): p. 443-448.

24. Roberto Todeschini, V.C., *Molecular Descriptors for Chemoinformatics*. 2010: Wiley-VCH Verlag GmbH & Co. KGaA. 1-38.

25. Willett, P., J.M. Barnard, and G.M. Downs, *Chemical similarity searching.* Journal of Chemical Information and Computer Sciences, 1998. **38**(6): p. 983-996.

26. Flower, D.R., *On the properties of bit string-based measures of chemical similarity.* Journal of Chemical Information and Computer Sciences, 1998. **38**(3): p. 379-386.

27. Bologa, C.G., et al., *Virtual and biomolecular screening converge on a selective agonist for GPR30.* Nature Chemical Biology, 2006. **2**(4): p. 207-212.

28. Stumpfe, D., et al., *Targeting Multifunctional Proteins by Virtual Screening: Structurally Diverse Cytohesin Inhibitors with Differentiated Biological Functions.* Acs Chemical Biology, 2010. **5**(9): p. 839-849.

29. Ijjaali, I., et al., *Ligand-based virtual screening to identify new T-type calcium channel blockers.* Channels, 2007. **1**(4): p. 300-4.

30. Hessler, G., et al., *Multiple-ligand-based virtual screening: methods and applications of the MTree approach.* Journal of medicinal chemistry, 2005. **48**(21): p. 6575-84.

31. Wermuth, C.G., *Pharmacophores: Historical Perspective and Viewpoint from a Medicinal Chemist*, in *Pharmacophores and Pharmacophore Searches*. 2006, Wiley-VCH Verlag GmbH & Co. KGaA. p. 1-13.

32. Wolber, G., et al., *Molecule-pharmacophore superpositioning and pattern matching in computational drug design.* Drug discovery today, 2008. **13**(1-2): p. 23-9.

33. Smellie, A., S.L. Teig, and P. Towbin, *Poling - Promoting Conformational Variation.* Journal of Computational Chemistry, 1995. **16**(2): p. 171-187.

34. Acharya, C., et al., *Recent advances in ligand-based drug design: relevance and utility of the conformationally sampled pharmacophore approach.* Current computer-aided drug design, 2011. **7**(1): p. 10-22.

35. Al-Sha'er, M.A. and M.O. Taha, *Elaborate ligand-based modeling reveals new nanomolar heat shock protein 90alpha inhibitors.* Journal of Chemical Information and Modeling, 2010. **50**(9): p. 1706-23.

36. Dixon, S.L., et al., *PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results.* J Comput Aided Mol Des, 2006. **20**(10-11): p. 647-71.

37. *Molecular Operating Environment (MOE)*, 2011, Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7.

38. Kurogi, Y. and O.F. Guner, *Pharmacophore modeling and three-dimensional database searching for drug design using catalyst.* Curr Med Chem, 2001. **8**(9): p. 1035-55.

39. *Catalyst*, 2002, Accelrys Inc.; San Diego, CA.

40. *LigandScout - advanced structure-based pharmacophore modeling.* 2012.

41. Chang, C. and P.W. Swaan, *Computational approaches to modeling drug transporters.* Eur J Pharm Sci, 2006. **27**(5): p. 411-24.

42. Schuster, D., et al., *Identification of chemically diverse, novel inhibitors of 17beta-hydroxysteroid dehydrogenase type 3 and 5 by pharmacophore-based virtual screening.* The Journal of steroid biochemistry and molecular biology, 2011. **125**(1-2): p. 148-61.

43. Noha, S.M., et al., *Discovery of a novel IKK-beta inhibitor by ligand-based virtual screening techniques.* Bioorganic & medicinal chemistry letters, 2011. **21**(1): p. 577-83.

44. Chiang, Y.K., et al., *Generation of ligand-based pharmacophore model and virtual screening for identification of novel tubulin inhibitors with potent anticancer activity.* Journal of medicinal chemistry, 2009. **52**(14): p. 4221-33.

45. Doddareddy, M.R., et al., *3D pharmacophore based virtual screening of T-type calcium channel blockers.* Bioorg Med Chem, 2007. **15**(2): p. 1091-105.

46. Annoura, H., et al., *Synthesis and biological evaluation of new 4-arylpiperidines and 4-aryl-4-piperidinols: dual Na(+) and Ca(2+) channel blockers with reduced affinity for dopamine D(2) receptors.* Bioorg Med Chem, 2002. **10**(2): p. 371-83.

47. Manetti, F., et al., *N-(thiazol-2-yl)-2-thiophene carboxamide derivatives as Abl inhibitors identified by a pharmacophore-based database screening of commercially available compounds.* Bioorganic & medicinal chemistry letters, 2008. **18**(15): p. 4328-31.

48. Lanier, M.C., et al., *Selection, synthesis, and structure-activity relationship of tetrahydropyrido[4,3-d]pyrimidine-2,4-diones as human GnRH receptor antagonists.* Bioorg Med Chem, 2007. **15**(16): p. 5590-603.

49. Roche, O. and R.M. Rodriguez Sarmiento, *A new class of histamine H3 receptor antagonists derived from ligand based design.* Bioorganic & medicinal chemistry letters, 2007. **17**(13): p. 3670-5.

50. Stahl, M., et al., *A validation study on the practical use of automated de novo design.* J Comput Aided Mol Des, 2002. **16**(7): p. 459-78.

51. Howells, L.M., et al., *Indole-3-carbinol inhibits protein kinase B/Akt and induces apoptosis in the human breast tumor cell line MDA MB468 but not in the nontumorigenic HBL100 line.* Mol Cancer Ther, 2002. **1**(13): p. 1161-72.

52. Li, Y., S.R. Chinni, and F.H. Sarkar, *Selective growth regulatory and pro-apoptotic effects of DIM is mediated by AKT and NF-kappaB pathways in prostate cancer cells.* Front Biosci, 2005. **10**: p. 236-43.

53. Chao, W.R., et al., *Computer-aided rational drug design: a novel agent (SR13668) designed to mimic the unique anticancer mechanisms of dietary indole-3-carbinol to block Akt signaling.* Journal of medicinal chemistry, 2007. **50**(15): p. 3412-5.

54. Reid, J.M., et al., *Phase 0 clinical chemoprevention trial of the Akt inhibitor SR13668.* Cancer prevention research, 2011. **4**(3): p. 347-53.

55. Sato, M., et al., *Novel HIV-1 integrase inhibitors derived from quinolone antibiotics.* Journal of Medicinal Chemistry, 2006. **49**(5): p. 1506-8.

56. Dayam, R., et al., *Quinolone 3-carboxylic acid pharmacophore: design of second generation HIV-1 integrase inhibitors.* Journal of medicinal chemistry, 2008. **51**(5): p. 1136-44.

57. Mugnaini, C., et al., *Toward novel HIV-1 integrase binding inhibitors: molecular modeling, synthesis, and biological studies.* Bioorg Med Chem Lett, 2007. **17**(19): p. 5370-3.

58. Noeske, T., et al., *Virtual screening for selective allosteric mGluR1 antagonists and structure-activity relationship investigations for coumarine derivatives.* Chemmedchem, 2007. **2**(12): p. 1763-73.

59. Bordi, F. and A. Ugolini, *Group I metabotropic glutamate receptors: implications for brain diseases.* Prog Neurobiol, 1999. **59**(1): p. 55-79.

60. Spooren, W., et al., *Insight into the function of Group I and Group II metabotropic glutamate (mGlu) receptors: behavioural characterization and implications for the treatment of CNS disorders.* Behav Pharmacol, 2003. **14**(4): p. 257-77.

61. Schneider, G., et al., *"Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening.* Angew Chem Int Ed Engl, 1999. **38**(19): p. 2894-2896.

62. Zhang, S., *Computer-aided drug discovery and development.* Methods in molecular biology, 2011. **716**: p. 23-38.

63. Hansch, C., *Citation Classic - Rho-Sigma-Pi-Analysis - a Method for the Correlation of Biological-Activity and Chemical-Structure.* Current Contents/Life Sciences, 1982(47): p. 18-18.

64. Free, S.M., Jr. and J.W. Wilson, *A Mathematical Contribution to Structure-Activity Studies.* Journal of Medicinal Chemistry, 1964. **7**: p. 395-9.

65. Tmej, C., et al., *A combined Hansch/Free-Wilson approach as predictive tool in QSAR studies on propafenone-type modulators of multidrug resistance.* Arch Pharm (Weinheim), 1998. **331**(7-8): p. 233-40.

66. Hemmer, M.C., V. Steinhauer, and J. Gasteiger, *Deriving the 3D structure of organic molecules from their infrared spectra.* Vibrational Spectroscopy, 1999. **19**(1): p. 151-164.

67. Schuur, J.H., P. Selzer, and J. Gasteiger, *The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity.* Journal of Chemical Information and Computer Sciences, 1996. **36**(2): p. 334-344.

68. Pearlman, R.S. and K.M. Smith, *Metric validation and the receptor-relevant subspace concept.* Journal of Chemical Information and Computer Sciences, 1999. **39**(1): p. 28-35.

69. Bravi, G., et al., *MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: A comparative 3D QSAR study in a series of steroids.* J Comput Aided Mol Des, 1997. **11**(1): p. 79-92.

70. Randic, M., *Molecular Profiles - Novel Geometry-Dependent Molecular Descriptors.* New Journal of Chemistry, 1995. **19**(7): p. 781-791.

71. Hong, H., et al., *Mold(2), molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics.* Journal of Chemical Information and Modeling, 2008. **48**(7): p. 1337-44.

72. Cramer, R.D., D.E. Patterson, and J.D. Bunce, *Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins.* Journal of the American Chemical Society, 1988. **110**(18): p. 5959-67.

73. Ekins, S., J. Mestres, and B. Testa, *In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling.* British journal of pharmacology, 2007. **152**(1): p. 9-20.

74. Kubinyi, H., G. Folkers, and Y.C. Martin, *3D QSAR in drug design*. Qdsar. 1998, Dordrecht ; Boston, Mass: Kluwer Academic. v. < 2- >.

75. Klebe, G., U. Abraham, and T. Mietzner, *Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity.* Journal of Medicinal Chemistry, 1994. **37**(24): p. 4130-46.

76. Hristozov, D.P., T.I. Oprea, and J. Gasteiger, *Virtual screening applications: a study of ligand-based methods and different structure representations in four different scenarios.* Journal of computer-aided molecular design, 2007. **21**(10-11): p. 617-40.

77. Wold, S., K. Esbensen, and P. Geladi, *Principal Component Analysis.* Chemometrics and Intelligent Laboratory Systems, 1987. **2**(1-3): p. 37-52.

78. Kubinyi, H., *QSAR and 3D QSAR in drug design .1. methodology.* Drug Discovery Today, 1997. **2**(11): p. 457-467.

79. Zheng, W. and A. Tropsha, *Novel variable selection quantitative structure--property relationship approach based on the k-nearest-neighbor principle.* J Chem Inf Comput Sci, 2000. **40**(1): p. 185-94.

80. Livingstone, D., *Artificial neural networks : methods and applications*. Methods in molecular biology,. 2008, Totowa, NJ: Humana Press. ix, 254 p.

81. Han, L.Y., et al., *Support vector machines approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness.* Drug Discovery Today, 2007. **12**(7-8): p. 304-13.

82. Casanola-Martin, G.M., et al., *Dragon method for finding novel tyrosinase inhibitors: Biosilico identification and experimental in vitro assays.* European journal of medicinal chemistry, 2007. **42**(11-12): p. 1370-81.

83. Gasparini, F., et al., *mGluR5 antagonists: discovery, characterization and drug development.* Curr Opin Drug Discov Devel, 2008. **11**(5): p. 655-65.

84. Conn, P.J., A. Christopoulos, and C.W. Lindsley, *Allosteric modulators of GPCRs: a novel approach for the treatment of CNS disorders.* Nat Rev Drug Discov, 2009. **8**(1): p. 41-54.

85. Mueller, R., et al., *Discovery of 2-(2-benzoxazoyl amino)-4-aryl-5-cyanopyrimidine as negative allosteric modulators (NAMs) of metabotropic glutamate receptor 5 (mGlu(5)): from an artificial neural network virtual screen to an in vivo tool compound.* ChemMedChem, 2012. **7**(3): p. 406-14.

86. Mueller, R., et al., *Identification of Metabotropic Glutamate Receptor Subtype 5 Potentiators Using Virtual High-Throughput Screening.* ACS Chem Neurosci, 2010. **1**(4): p. 288-305.

87. Rodriguez, A.L., et al., *Discovery of novel allosteric modulators of metabotropic glutamate receptor subtype 5 reveals chemical and functional diversity and in vivo activity in rat behavioral models of anxiolytic and antipsychotic activity.* Mol Pharmacol, 2010. **78**(6): p. 1105-23.

88. Feher, M. and J.M. Schmidt, *Property distributions: Differences between drugs, natural products, and molecules from combinatorial chemistry.* Journal of Chemical Information and Computer Sciences, 2003. **43**(1): p. 218-227.

89. Bohacek, R.S., C. McMartin, and W.C. Guida, *The art and practice of structure-based drug design: A molecular modeling perspective.* Medicinal Research Reviews, 1996. **16**(1): p. 3-50.

90. Schmidt, J.M., et al., *De Novo Design, Synthesis, and Evaluation of Novel Nonsteroidal Phenanthrene Ligands for the Estrogen Receptor.* Journal of Medicinal Chemistry, 2003. **46**(8): p. 1408-1418.

91. Lloyd, D.G., et al., *Scaffold Hopping in De Novo Design. Ligand Generation in the Absence of Receptor Information.* Journal of Medicinal Chemistry, 2003. **47**(3): p. 493-496.

92. Waszkowycz, B., et al., *PRO_LIGAND: An Approach to de Novo Molecular Design. 2. Design of Novel Molecules from Molecular Field Analysis (MFA) Models and Pharmacophores.* Journal of Medicinal Chemistry, 1994. **37**(23): p. 3994-4002.

93. Schneider, G., et al., *De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks.* J Comput Aided Mol Des, 2000. **14**(5): p. 487-94.

94. Rogers-Evans, M., et al., *Identification of novel cannabinoid receptor ligands via evolutionary de novo design and rapid parallel synthesis.* Qsar & Combinatorial Science, 2004. **23**(6): p. 426-430.

95. Brown, N., et al., *A Graph-Based Genetic Algorithm and Its Application to the Multiobjective Evolution of Median Molecules.* Journal of Chemical Information and Computer Sciences, 2004. **44**(3): p. 1079-1087.

96. Feher, M., et al., *The use of ligand-based de novo design for scaffold hopping and sidechain optimization: two case studies.* Bioorganic & medicinal chemistry, 2008. **16**(1): p. 422-7.

97. Golla, S., et al., *Virtual design of chemical penetration enhancers for transdermal drug delivery.* Chemical biology & drug design, 2012. **79**(4): p. 478-87.

98. Baumforth, K.R., et al., *Induction of autotaxin by the Epstein-Barr virus promotes the growth and survival of Hodgkin lymphoma cells.* Blood, 2005. **106**(6): p. 2138-46.

99. Euer, N., et al., *Identification of genes associated with metastasis of mammary carcinoma in metastatic versus non-metastatic cell lines.* Anticancer Res, 2002. **22**(2A): p. 733-40.

100. Kawagoe, H., et al., *Expression and transcriptional regulation of the PD-Ialpha/autotaxin gene in neuroblastoma.* Cancer Res, 1997. **57**(12): p. 2516-21.

101. Boucher, J., et al., *Potential involvement of adipocyte insulin resistance in obesity-associated up-regulation of adipocyte lysophospholipase D/autotaxin expression.* Diabetologia, 2005. **48**(3): p. 569-77.

102. Umemura, K., et al., *Autotaxin expression is enhanced in frontal cortex of Alzheimer-type dementia patients.* Neurosci Lett, 2006. **400**(1-2): p. 97-100.

103. Inoue, M., et al., *Simultaneous stimulation of spinal NK1 and NMDA receptors produces LPC which undergoes ATX-mediated conversion to LPA, an initiator of neuropathic pain.* J Neurochem, 2008. **107**(6): p. 1556-65.

104. Hoeglund, A.B., et al., *Optimization of a pipemidic acid autotaxin inhibitor.* Journal of medicinal chemistry, 2010. **53**(3): p. 1056-66.

105. Kurup, A., *C-QSAR: a database of 18,000 QSARs and associated biological and physical data.* J Comput Aided Mol Des, 2003. **17**(2-4): p. 187-96.

106. Kim, K.H., *Outliers in SAR and QSAR: 2. Is a flexible binding site a possible source of outliers?* J Comput Aided Mol Des, 2007. **21**(8): p. 421-35.

107. Prelog, V. and G. Helmchen, *Basic Principles of the Cip-System and Proposals for a Revision.* Angewandte Chemie-International Edition in English, 1982. **21**(8): p. 567-583.

108. Schiffman, S.S., T.B. Clark, 3rd, and J. Gagnon, *Influence of chirality of amino acids on the growth of perceived taste intensity with concentration.* Physiol Behav, 1982. **28**(3): p. 457-65.

109. Pai, V. and N. Pai, *Recent advances in chirally pure proton pump inhibitors.* J Indian Med Assoc, 2007. **105**(8): p. 469-70, 472, 474.

110. Mehvar, R. and D.R. Brocks, *Stereospecific pharmacokinetics and pharmacodynamics of beta-adrenergic blockers in humans.* J Pharm Pharm Sci, 2001. **4**(2): p. 185-200.

111. Gurjar, M.K., *The future lies in chiral purity: a perspective.* J Indian Med Assoc, 2007. **105**(4): p. 177-8.

112. Francotte, E. and W. Lindner, *Chirality in drug research*. Methods and principles in medicinal chemistry. 2006, Weinheim: Wiley-VCH. xix, 351 p.

113. Beroza, P. and M.J. Suto, *Designing chiral libraries for drug discovery.* Drug Discovery Today, 2000. **5**(8): p. 364-372.

114. Murakami, H., *From racemates to single enantiomers - Chiral synthetic drugs over the last 20 years.* Novel Optical Resolution Technologies, 2007. **269**: p. 273-299.

115. Golbraikh, A., D. Bonchev, and A. Tropsha, *Novel chirality descriptors derived from molecular topology.* Journal of Chemical Information and Computer Sciences, 2001. **41**(1): p. 147-158.

116. Yang, C.S. and C.L. Zhong, *Chirality factors and their application to QSAR studies of chiral molecules.* Qsar & Combinatorial Science, 2005. **24**(9): p. 1047-1055.

117. Brown, J.B., et al., *Compound Analysis Via Graph Kernels Incorporating Chirality.* Journal of Bioinformatics and Computational Biology, 2010. **8**: p. 63-81.

118. Lukovits, I. and W. Linert, *A topological account of chirality.* Journal of Chemical Information and Computer Sciences, 2001. **41**(6): p. 1517-1520.

119. Marrero-Ponce, Y. and J.A. Castillo-Garit, *3D-chiral atom, atom-type, and total non-stochastic and stochastic molecular linear indices and their applications to central chirality codification.* Journal of Computer-Aided Molecular Design, 2005. **19**(6): p. 369-383.

120. Del Rio, A., *Exploring enantioselective molecular recognition mechanisms with chemoinformatic techniques.* Journal of Separation Science, 2009. **32**(10): p. 1566-1584.

121. Benigni, R., et al., *Deriving a quantitative chirality measure from molecular similarity indices.* J Med Chem, 2000. **43**(20): p. 3699-3703.

122. Zabrodsky, H., S. Peleg, and D. Avnir, *Continuous Symmetry Measures.* Journal of the American Chemical Society, 1992. **114**(20): p. 7843-7851.

123. Aires-de-Sousa, J. and J. Gasteiger, *New description of molecular chirality and its application to the prediction of the preferred enantiomer in stereoselective reactions.* Journal of Chemical Information and Computer Sciences, 2001. **41**(2): p. 369-375.

124. Aires-de-Sousa, J. and J. Gasteiger, *Prediction of enantiomeric selectivity in chromatography - Application of conformation-dependent and conformation-independent descriptors of molecular chirality.* Journal of Molecular Graphics & Modelling, 2002. **20**(5): p. 373-388.

125. Aires-De-Sousa, J., et al., *Chirality codes and molecular structure.* Journal of Chemical Information and Computer Sciences, 2004. **44**(3): p. 831-836.

126. Verma, J., V.M. Khedkar, and E.C. Coutinho, *3D-QSAR in Drug Design - A Review.* Current Topics in Medicinal Chemistry, 2010. **10**(1): p. 95-115.

127. Cramer, R.D., D.E. Patterson, and J.D. Bunce, *Comparative Molecular-Field Analysis (Comfa) .1. Effect of Shape on Binding of Steroids to Carrier Proteins.* Journal of the American Chemical Society, 1988. **110**(18): p. 5959-5967.

128. Silverman, D.B., *The thirty-one benchmark steroids revisited: Comparative molecular moment analysis (CoMMA) with principal component regression.* Quantitative Structure-Activity Relationships, 2000. **19**(3): p. 237-246.

129. Robert, D., L. Amat, and R. Carbo-Dorca, *Three-dimensional quantitative structure-activity relationships from tuned molecular quantum similarity measures: Prediction of the corticosteroid-binding globulin binding affinity for a steroid family.* Journal of Chemical Information and Computer Sciences, 1999. **39**(2): p. 333-344.

130. So, S.S. and M. Karplus, *Three-dimensional quantitative structure-activity relationships from molecular similarity matrices and genetic neural networks. 1. Method and validations.* J Med Chem, 1997. **40**(26): p. 4347-4359.

131. Klebe, G., U. Abraham, and T. Mietzner, *Molecular Similarity Indexes in a Comparative-Analysis (Comsia) of Drug Molecules to Correlate and Predict Their Biological-Activity.* J Med Chem, 1994. **37**(24): p. 4130-4146.

132. Maw, H.H. and L.H. Hall, *E-state modeling of corticosteroids binding affinity validation of model for small data set.* Journal of Chemical Information and Computer Sciences, 2001. **41**(5): p. 1248-1254.

133. Liu, S.S., C.S. Yin, and L.S. Wang, *Combined MEDV-GA-MLR method for QSAR of three panels of steroids, dipeptides, and COX-2 inhibitors.* Journal of Chemical Information and Computer Sciences, 2002. **42**(3): p. 749-756.

134. Besalu, E., et al., *Molecular quantum similarity and the fundamentals of QSAR.* Accounts of Chemical Research, 2002. **35**(5): p. 289-295.

135. Gasteiger, J. and M. Marsili, *New Model for Calculating Atomic Charges in Molecules.* Tetrahedron Letters, 1978(34): p. 3181-3184.

136. Gasteiger, J. and M. Marsili, *Iterative Partial Equalization of Orbital Electronegativity - a Rapid Access to Atomic Charges.* Tetrahedron, 1980. **36**(22): p. 3219-3228.

137. Guillen, M.D. and J. Gasteiger, *Extension of the Method of Iterative Partial Equalization of Orbital Electronegativity to Small Ring-Systems.* Tetrahedron, 1983. **39**(8): p. 1331-1335.

138. Bauerschmidt, S. and J. Gasteiger, *Overcoming the limitations of a connection table description: A universal representation of chemical species.* Journal of Chemical Information and Computer Sciences, 1997. **37**(4): p. 705-714.

139. Streitwieser, A., *Molecular orbital theory for organic chemists*. 1961, New York,: Wiley. 489 p.

140. Gasteiger, J. and H. Saller, *Calculation of the Charge-Distribution in Conjugated Systems by a Quantification of the Resonance Concept.* Angewandte Chemie-International Edition in English, 1985. **24**(8): p. 687-689.

141. Gilson, M.K., H.S.R. Gilson, and M.J. Potter, *Fast assignment of accurate partial atomic charges: An electronegativity equalization method that accounts for alternate resonance forms.* Journal of Chemical Information and Computer Sciences, 2003. **43**(6): p. 1982-1997.

142. Gasteiger, J. and M.G. Hutchings, *New Empirical-Models of Substituent Polarizability and Their Application to Stabilization Effects in Positively Charged Species.* Tetrahedron Letters, 1983. **24**(25): p. 2537-2540.

143. Gasteiger, J. and M.G. Hutchings, *Quantitative Models of Gas-Phase Proton-Transfer Reactions Involving Alcohols, Ethers, and Their Thio Analogs - Correlation Analyses Based on Residual Electronegativity and Effective Polarizability.* Journal of the American Chemical Society, 1984. **106**(22): p. 6489-6495.

144. Miller, K.J., *Additivity Methods in Molecular Polarizability.* Journal of the American Chemical Society, 1990. **112**(23): p. 8533-8542.

145. Wang, R.X., Y. Gao, and L.H. Lai, *Calculating partition coefficient by atom-additive method.* Perspectives in Drug Discovery and Design, 2000. **19**(1): p. 47-66.